THE UNIVERSITY
of EDINBURGH

High-resolution Fourier transform ion cyclotron resonance mass spectrometry and nuclear magnetic resonance spectroscopy of humic substances

John Blackburn

Thesis submitted for the degree of Doctor of Philosophy

School of Chemistry

University of Edinburgh

2017

Abstract

Humic substances (HS) are described as a complex mixture of organic molecules formed by incomplete decomposition of plant, animal and microbial matter. They are found in soil, water and air and have many environmental roles, *e.g.* water retention and metal ion binding in soil. Despite their importance, the molecular composition of HS is poorly understood. This is mostly because of an inability to separate individual molecules from these complex mixtures and then characterise them by standard analytical methods such as NMR and MS. In order improve the understanding of these important mixtures I have studied them using a high-resolution analytical method, Fourier transform ion-cyclotron resonance mass spectrometry (FTICR MS).

Initial efforts focussed on testing the, fast, automated data analysis of the large data sets produced. Two pieces of software were compared and the reliability of the formulae assigned by these was critically evaluated. This confident formula assignment was then applied to study the consequences of different ionisation and instrumental parameters on the mass spectra obtained. The use of laser desorption/ionisation (LDI) without the need to employ a matrix required in matrix assisted laser desorption/ionisation (MALDI) was explored. A comparison of LDI and electrospray ionisation (ESI) FTICR MS of natural organic matter samples showed that these methods ionise complementary sets of compounds. The LDI ionised compounds were characterised as aromatics or condensed aromatics and compounds belonging to lower oxygen classes (maximum number at O8), while ESI ionised higher oxygen classes (maximum number at O16) with a vast majority of compounds classified as aliphatic based on their modified aromaticity index. MALDI and LDI spectra produced very similar data with over 90% matching formulas implying that fragmentation is not caused by LDI, as taught previously. My work showed that to maximize the coverage by FTICR MS of the molecular space occupied by these complex mixtures, multiple ionization methods must be used. As a particularly convenient and readily deployable ionization technique, LDI should be included in standard analytical protocols for FTICR MS analysis of NOM.

I have explored different parameters and experimental settings to obtain a fuller coverage of the molecular space of NOM, this showed that different experimental

conditions enhance peak intensities in different *m/z* regions of the FTICR MS spectra and that information can be obtained outside of the narrow 200-700 *m/z* window.

To gain chemical and structural information about humic substances beyond what is currently known, experiments aimed to label HS using different isotopes and at specific sites were developed and tested. Two methylation reactions were of particular interest. A methylation that selectively targeted carboxylic acid groups and incorporated deuterium in the form of CD_3 groups. An international standard, Suwannee River fulvic acid, was methylated and analysed by high-resolution mass spectrometry in order to gain information on the number and distribution carboxylic acid groups. This proved challenging due to the reactivity of the unknown molecules being difficult to determine in advance. Additionally, the peak separation being reduced to as low as 1.5 mDa pushed the instrument resolution and assignment confidence to their limits.

The second methylation method explored used $^{13}CH_3I$, a nonselective agent reacting with any labile proton, particularly attaching $^{13}CH_3$ groups to carboxylic, phenolic and alcoholic OH groups. I prepared a methylated sample of fulvic acid from a Red Moss raised bog (Balerno, near Edinburgh) ready for analyses by high field NMR. This investigation yielded structures of a number of phenolic compounds for the first time by NMR.

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or processional qualification except as specified.

Parts of this work have been published in:

Bell, N.G.A., Michalchuk, A.A.L., **Blackburn**, **J**.**W**.**T**., Graham, M.C. and Uhrín, D., *Isotope-Filtered 4D NMR Spectroscopy for Structure Determination of Humic Substances.* Angewandte Chemie (International Ed. in English), 2015. **54**(29): p. 8382-8385.

Kew, W., **Blackburn**, **J**.**W**.**T**., Clarke, D.J. and Uhrín, D., *Interactive van Krevelen diagrams – Advanced visualisation of mass spectrometry data of complex mixtures.* Rapid Communications in Mass Spectrometry, 2017. **31**(7): p. 658-662.

Blackburn, J.W.T., Kew, W., Graham, M.C. and Uhrín, D., *Laser Desorption/Ionization Coupled to FTICR Mass Spectrometry for Studies of Natural Organic Matter.* Analytical Chemistry, 2017. **89**(8): p. 4382-4386.

Kew, W., **Blackburn, J.W.T**, and Uhrín, D., *Response to comment on "Laser Desorption/Ionization Coupled to FTICR Mass Spectrometry for Studies of Natural Organic Matter".* Analytical Chemistry, 2018. **90**(9): p. 5968-5971.

John W. T. Blackburn. May 2018

Acknowledgments

I would like to thank my supervisors Prof. Dušan Uhrín and Dr Margaret Graham for providing me with the opportunity to study such an interesting and challenging field. Additionally, I would not have been able to achieve anything that is shown here without the excellent technical support of Dr Logan Mackay, Mr Juraj Bella and Dr Lorna Murray, to you I am eternally grateful. Thank you to my project students Xu, Yadi and Leyla for all their diligent work in the lab.

The Geo-boys of lab 293, Andrew, Michael and Gavin, you three and the combination of digger chat, fieldwork and outdoor shenanigans made sure I never truly left my countryside roots behind. In remembrance of Heli-Gavin, may he deflate in peace.

Thank you to the whole NMR group, who have provided endless entertainment, and biscuits, over the years. Special thanks go to Natalia, Hannah, Will and Adam; who have somehow managed to keep me smiling, exercising, sleeping (somewhat) and eating foods other than potato waffles.

The man, the myth, the legend. Dr Haris Panagos, I can't begin to praise you enough, you've been a motivator, a cynic, an accomplice, an inspiration but most importantly a friend. You've taught me and helped me more than you know, and made the ride an incredibly fun one.

Leigh, you have always been there to keep me grounded, to keep me sane, no matter how fried my brain became. It's been 15 years since you took on such a foolhardy role of being my friend and I am who I've become, in part, thanks to you, it's your fault.

To all my friends and family who have supported me throughout, I thank you all. I am especially thankful for all the support of my wonderful siblings, I'm lucky to be your brother and owe you all at least 1 beer (coupon valid until 01/09/2018).

Finally, I would like to thank Alice Extance, who gave me the courage to follow my dreams and allowed me to take a path that has brought me many new experiences, challenged me in ways I couldn't predict and allowed me to grow immeasurably.

Contents

Abbreviations

AI	Aromaticity index
APCI	Atmospheric pressure chemical ionisation
APPI	Atmospheric pressure photoionisation
CID	Collision induced dissociation
COSY	Correlated spectroscopy
CRAM	Carboxylic-rich alicyclic molecules
CRM	Charge residue model
DBE	Double bond equivalents
DOM	Dissolved organic matter
EEM	Excitation-emission matrix spectroscopy
ESI	Electrospray ionisation
FA	Fulvic acid
FTICR MS	Fourier transform ion-cyclotron resonance mass spectrometry
HA	Humic acid
HS	Humic substances
IAT	Ion accumulation time
IEM	Ion evaporation model
IHSS	International Humic Substances Society
IRMPD	Infrared multiphoton dissociation
KDE	Kernel density estimation
KM	Kendrick mass
KMD	Kendrick mass defect
LDI	Laser desorption/ionisation
MALDI	Matrix-assisted laser desorption/ionisation
MDLT	Material derived from linear terpenoids
NMR	Nuclear magnetic resonance
NOM	Natural organic matter

RMFA	Red Moss fulvic acid
RMHA	Red Moss humic acid
SNR	Signal-to-noise ratio
SOM	Soil organic matter
SRFA	Suwannee river fulvic acid
SWIM	Stored-waveform ion modulation
TOCSY	Total correlation spectroscopy

Chapter 1: Introduction

1.1 Organic matter in the environment

Despite its mundane appearance, soil is one of the most importance substances on our planet. It is home to a wide variety of plants, animals and microbes. All of which depend on it greatly, even if indirectly. Following any food chain, even our own, will eventually lead to plants that without soil, or more importantly the chemical substances held within the soil, could not grow.

Soil is the anchor for most plant life, it has the ability to maintain a sustainable ecosystem and not dry up of resources through the aid of a host of animals, microbes and fungi. It would be impossible for an area of ground to sustain continual plant growth successfully for the foreseeable future without some means of returning necessary materials to the soil. Fungi break down the large plant material (*e.g.* bark, branches and leaves) into smaller compounds and microbes continue the process to fully break down the plant matter (humification). Some animals like earth worms rapidly improve this process by physically moving the litter deeper underground into their burrows and by feeding on this matter. The same microbes that exist in the soil exist in the gut of the earth worms. However, the environment within the worm is much more suitable for the microbes in terms of variables like temperature and pH. This allows the microbes to perform up to 1000 times faster when breaking down the plant material. The result is a continually cycling system of compounds that ultimately allows sustainable, long term growth; creating a symbiosis between the plant and animal kingdoms.

1.1.1 Natural Organic Matter

An approximation of the total carbon in the environment is 350 x 10^{15} g,[1] from this organic matter within soil accounts for three times the amount of organic carbon in comparison to that of the atmosphere and living plants combined.[2] Soil therefore plays an important role in the biogeochemical processes that occur within the environment due to being such a major store of organic carbon.[3]

Natural Organic Matter (NOM) describes the organic matter found in soil and water systems. These systems can be separately described as soil organic matter (SOM) and dissolved organic matter (DOM). DOM by classification is the organic matter in water systems that is smaller than 0.45 μm in diameter, with anything larger being described as particulate organic matter.[4] In both cases the organic matter can be broken into two major fractions, the nonhumified organic matter and the humified organic matter.

1.1.2 Nonhumified organic matter

Nonhumified organic matter describes a group of compounds detectable within soil systems that are present in plant organisms. The main compound classes are carbohydrates, proteins and amino acids, lipids, nucleic acids and lignin.[5]

The carbohydrates present in soil are in the form of polysaccharides, such as cellulose, and monosaccharides, such as glucose and galactose. Although, these free monosaccharides account for an insignificantly small faction of the total carbohydrate level is a soil sample. Carbohydrate levels in soil follow organic matter trends, with matching high and low concentration horizons.[6] In aerobic conditions these would

eventually decompose to CO_2 and water or in anaerobic conditions to CO_2, CH_4 and C_2H_5OH.

Proteins, amino acids and their degradation products act as a major source of nitrogen within soil. Due to the ability of amino acids to act as zwitterions they can be anionic, neutral or cationic depending on the pH of the soil and so can be found in any of these forms. Proteins are increasingly thought of as being a major reason for the stability of organic nitrogen in SOM. Either by the protein being released during cell death and performing functional roles within the soil, or by proteinases of microorganisms hydrolysing the peptide bond to release usable peptides and amino acids.[7] Once in the form of amino acids, a few degradation routes are possible depending on the immediate environment via dehydrogenases and oxidases. In aerobic conditions CO_2, NH_3 and RCOOH are formed. In anaerobic conditions with hydrogen present, deamination can occur to give NH_3 and RCH_2COOH or with decarboxylation amines and CO can be produced.

Lipids are in the form of a heterogeneous mixture of fatty acids, wax and oils. The most basic product of the degradation from these molecules is glycerol ($C_3H_8O_3$). Glycerol forms lipids by binding to fatty acids as mono-, di- and triglycerides. Lipids can be described as; simple lipids, such as natural plant oils; compound lipids, which are a mixtures of fatty acids with phosphorous containing organic compounds; and derived lipids, which contain alcohols and sterols. The most abundant type of lipid within soils are phospholipids[8], such as terpenoids, as they are common natural products. The lipid mixtures undergo hydrolysis via saponification; this breaks down the lipids into their respective building blocks.

Nucleic acids come from the DNA and RNA of plants and microbes. They are phosphorous and nitrogen containing compounds that start as DNA/RNA and break down initially into their nucleic acid monomers. DNA is typically quite robust and long-lived in soil, especially in animal remains.

Lignin is a major component of the nonhumified organic matter. It is a highly aromatic polymer created from $C_xH_yO_z$ monomers via lignification. It can take on a wide variety of forms depending on the monomers incorporated. The three most common monolignols are *p*-coumaryl alcohol, coniferyl alcohol and sinapyl alcohol (Figure 1.1).[9] In plants the lignin packs the space between cellulose bundles and hemicellulose, which is why it needs to have many different structural forms, so it can fill this space as efficiently as possible. This crosslinking of the cellulose and hemicellulose in higher plants strengthens it against mechanical and chemical decomposition.[10] Lignin can be broken down by a variety of fungi. White rot in aerated conditions will break lignin down into CO_2 and water, brown rot will remove the $-OCH_3$ groups from the lignin and black rot will break up the polymer structure of lignin when in wet, low oxygen soils. If fungi are unable to break lignin down due to the soil conditions being unsuitable, the concentration of lignin can build up. This is due to lignin being highly resistant to microbial decomposition. If the lignin concentration within the soil reaches a high enough level, this can in the very long term lead to the formation of peat.[5]

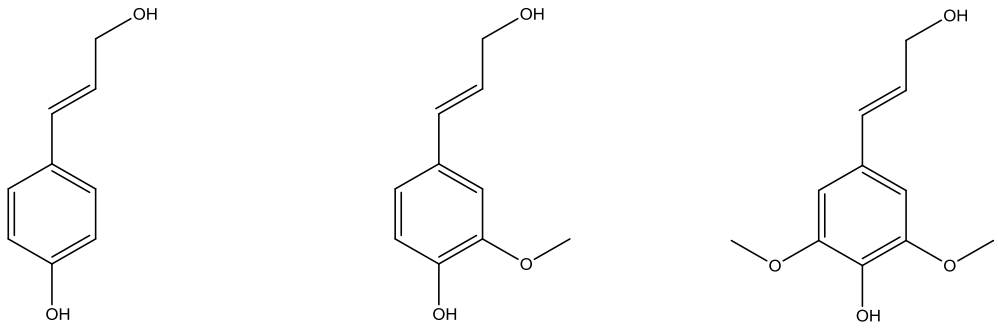


Figure 1.1 Structures of *p*-coumaryl alcohol, coniferyl alcohol and sinapyl alcohol (left to right)

1.1.3 Humified organic matter

Plant material on the surface of the soil is described as litter and is degraded by microbes, this degradation is known as humification. The result is the humified organic matter in the soil. These substances, which come from biopolymers in varied states of decay, don't fall into any discrete class of organic matter. Therefore they are termed humic substances (HS).[11, 12] Originally, HS were assumed to be large organic polymers. This assumption was based on the fact that many of these properties, including optical properties, exchange acidities, electrophoretic properties and MW characteristics were compatible with large molecules. These observables varied depending on the type and source of soil being examined.[13] It was theorised that HS contained a huge range of chemicals, containing a variety of chemical structures, which ranged from masses in the hundreds, to masses in the hundreds of thousands.[14] This was later challenged by Piccolo (2001) whom, through the use of low pressure size exclusion chromatography, suggested that HS were in fact a heterogeneous mixture of molecules with relatively much smaller masses than previously thought.

The small molecules are held in a supramolecular framework by van der Waals, π-π and CH-π interactions, which mimics a larger polymeric molecule when examined through most methods. Today, HS are viewed as a super complex mixture of small organic molecules held together by numerous intramolecular interactions.[15]

Although the exact chemical structures of HS are unknown, the biogeochemical effects have been more widely explored and identified. One role of HS is to bind cationic metals (*e.g.* Al(III), Fe(II) and Cu(II)),[16] which in turn effects the bioavailability, the availability of ions for plant and microbial uptake, of useful/nutrient ions[17] or aids in making potentially toxic ions insoluble.[12] On a longer geological timescale HS are thought to affect a multitude of processes. Some of these processed include; the weathering of rocks into soils; reducing the severity of mechanical soil erosion, such as rainfall impacting with the topsoil; stabilising atmosphere levels of O_2 caused by changing climates and forming precursors for fuel (coal/oil) production.[2, 17] HS also modulate surface temperature and provide stability to larger scale structures (*e.g.* slopes, verges).[2, 12, 17] When considering DOM, HS can be seen to control carbon and nitrogen cycles in aquatic environments, the quality of water in these environments, influence/undergo redox chemistry and are important in microbial processes. These effects are produced by DOM playing a significant role in the nutrition cycle at a microbial level, the decaying microbes and plants release DOM that other microbes and plants take up and use, allowing for the plant and microbial species to co-exist and co-evolve.[2, 18, 19]

If a more precise and reliable identification of structures could be obtained it would allow deeper understanding of the humification process. Also, it would allow temporal and spacial comparisons to be made between different HS samples. HS are known to

have an impact upon many geological processes. HS act as a long term storage of organic carbon in soils and it has been theorised that one possible cause of increased CO_2 release into the atmosphere is due to increased soil disturbance.[20] Carbon sequestration in soils is understood to be a major contributor to reducing atmospheric CO_2 levels, a variety of methods are used to attempt to reduce HS loss from the most exposed top-soils via erosion, especially for agricultural land.[21] Work has been performed with the goal of increasing the amount of organic carbon stored in soils in order to tackle current climate change issues.[22, 23] The most interesting functionality, from a chemistry perspective, in HS are carboxylic acid groups, these are known to bind and chelate, metal ions. By interacting with the metal ions HS stores and releases them based on changes in soil pH, oxygen content and water content.[24] HS have been shown to act in a similar manner as fertiliser by increasing plant uptake of preferential atoms like nitrogen and potassium. As opposed to fertilisers, HS are natural compounds found in the soil and would not cause as sever eutrophication, or similar negatives typically associated with fertilisers.[25-27] The concept of using HS as a fertiliser in conjugation with nitrogen fixing bacteria has also been tested and shown to have positive effects on plant growth.[28, 29] Justified by the functions of HS discussed here, and many additional functions not discussed here, combined with the fact that, for most cases, the mechanisms in which HS performs these functions is only speculated upon, more information on their structures is required.

1.1.4 Operational fractions of HS

HS can be separated into three operational fractions based on the solubility of different parts of the heterogeneous mix in different pH environments. These fractions are fulvic

acid (FA), humic acid (HA) and humin. FA is soluble over the entire pH range from acid to alkaline. HA is insoluble in acidic conditions and will precipitate out of solution that has been lowered to a pH below 2. Humin is insoluble in both strong acid and strong alkaline conditions and therefore cannot be solubilised in aqueous solutions.[30] It is customary to study individual operational fractions of HS rather than dealing with the entire NOM. This is due to HS being highly complex, these operational fractions offer a way to clearly break the mixture down into less complex sub-mixtures. The fractions are therefore easier to study than the original HS. The danger of this approach is that it only displays a subset of the sample, which makes it difficult to perform a comprehensive analysis.

In order to better understand these systems, I attempt to analyse them using high-resolution analytical techniques, nuclear magnetic resonance spectroscopy (NMR) and Fourier transform ion-cyclotron resonance mass spectrometry (FTICR MS); these techniques are introduced here.

1.2 Nuclear Magnetic Resonance spectroscopy

1.2.1 Principles of NMR

NMR is an analytical method that provides information on the chemical environment of selected nuclei by measuring their resonance frequency. When a diamagnetic nucleus (most elements have at least one diamagnetic isotope) is introduced into a magnetic field, the magnetic moment is forced into several allowed orientations. An example of this would be the hydrogen atom, which has two allowed orientations. The magnetic moment can point in the opposite direction to the field or in the same

direction as the field (which is the lowest energy state). The difference in energy between these two states, the resonance condition (ΔE), is dependent on the nucleus, the strength of the field and slightly on the environment of the nucleus. This small dependence on the environment is what allows the chemical environment of each nucleus to be examined. By applying electromagnetic radiation that flips the nuclei from the lower energy state to the higher state, following the equation $\Delta E = h\nu$ where h = Planck's constant and ν = frequency of radiation applied, ΔE can be measured.

A magnetic moment, μ, is produced by a combination of angular momentum ($\vec{P}$) and charge. This can be written as as $\vec{\mu} = \gamma\vec{P}$. γ is the magnetogyric ratio which describes the strength of the magnetic property for the given nucleus and is constant for all like nuclei.[31] This means that all magnetic nuclei have an angular momentum, or spin. The spin has a magnitude and a quantum number associated with it. The quantum number is denoted as *I* and can equal 0, 1/2, 1, 3/2 etc. The magnitude of spin (in units of ħ) is dependent on *I* and can be calculated using Equation 1.[32]

$$Magnitude\ of\ spin = (I(I+1))^{\frac{1}{2}}\ \hbar \qquad (1)$$

For the spin quantum number, *I*, the number of allowed configurations, or spin states, is equal to $2I + 1$. So for ^{1}H with a spin quantum number of ½ there are two allowed spin states (Figure 1.2). The resonance condition is the energy between these two spin states.

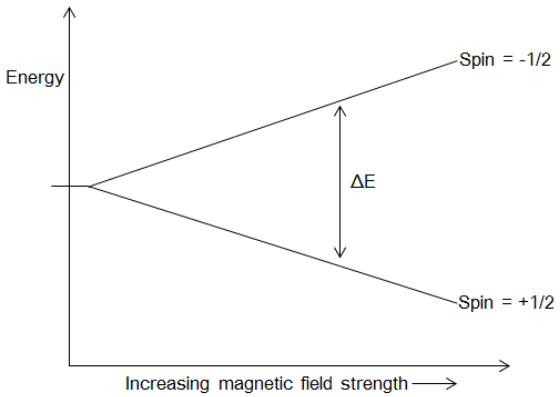


Figure 1.2 Spatial representation of the spin states separating in energy within a magnetic field

As previously stated, most nuclei have at least one diamagnetic isotope. However, this is not always the most abundant isotope. It is for nuclei like ^{1}H and ^{19}F, both of which have $I = ½$ and have natural abundances of 99.985 % and 100 % respectively. For carbon on the other hand its most abundant isotope ^{12}C has a spin quantum number of 0 and it is therefore non-magnetic. ^{13}C has an $I = ½$, similar to hydrogen, but only has a natural abundance of 1.1 %. This ultimately leads to ^{13}C NMR being much less sensitive compared to ^{1}H NMR.

Every equivalent nucleus of a particular element should resonate at the same frequency. However, the environment around the nucleus has a small effect on this frequency, causing nuclei of the same element in different environments to differentiate. This is called chemical shift and is what allows NMR users to analyse the environment of different nuclei within the molecule. Chemical shift arises because the external field of the spectrometer (B_0) induces movement in the electrons associated with the nucleus. This movement forms a magnetic field in the opposing

direction to the external field (B'). The result is that the nucleus feels a slightly smaller external field (B) due to the shielding effect of its electrons (Equation 2). The strength of the field applied to a nucleus, B, affects the value of ΔE and thus the resonance frequency. An example of this is shown in Figure 1.3.

$$B = B_0 - B' \tag{2}$$

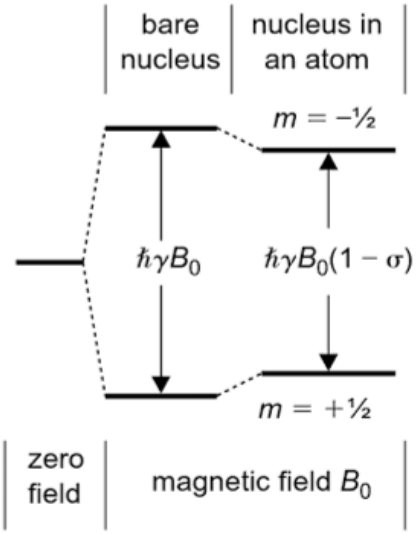


Figure 1.3 Energy level diagram for a nucleus in a magnetic field (B_0) and an in an example atom.[32]

B' is dependent on the size of B_0 and the electron density on the nucleus. When the electron density on the nucleus alters, the value of B' alters. As a consequence, so does B, ΔE, the resonance frequency and chemical shift. The environment around the nucleus affects the electron density on the nucleus by either withdrawing from or donating to the electron density of the nucleus. By withdrawing electron density away,

the value of B' is reduced, increasing the value of B, ΔE and the chemical shift. The opposite argument works for electron donating groups. This is described as de-shielding/shielding respectively. For example, benzene protons resonate at a chemical shift of 7.28 ppm. However, if electron donating/withdrawing groups are close enough to shield/de-shield the nucleus of the aromatic proton, they could have a chemical shift anywhere from 6.5 – 8.5 ppm. Equally, an aliphatic methyl proton could have a chemical shift from 0.8 – 1 ppm. So not only can different functionalities be differentiated from one another but also differences in the immediate environments of similar nuclei can be identified.

Another way of looking into the immediate environment of a specific nuclei is coupling. Coupling occurs when another non-equivalent magnetic nucleus is within 2 – 3 (sometimes 4) bonds of the observed nucleus. The second magnetic nucleus can exist in any of its magnetic quantum number (m) states (*e.g.* for $I = ½$, m = + ½ or - ½ and for $I = 1$, m = 1, 0, -1). The nucleus is the source of a small local magnetic field which can align with or oppose the external field. This affects B in the same manner as described in the above section and has the same consequences, although at a much lesser scale. However, as all states are potentially equally possible for the second nucleus rather than just shifting the value of ΔE, it is split into multiple values (depending on the number of states possible). For example, if a hydrogen nucleus (H_1) is close to a non-equivalent hydrogen nucleus (H_2) both nuclei can exist in either the + ½ state (H+) or the – ½ state (H-). This leads to four possible orientations: H_1+/H_2+, H_1+/H_2-, H_1-/H_2+ and H_1-/H_2-. So rather than have one ΔE value between H_1+ and H_1- there are now two ΔE values. One slightly larger ΔE between H_1+/H_2- and H_1-/H_2- and one slightly smaller between H_1+/H_2+ and H_1-/H_2+. The result is two frequencies of

equal probability an equal amount above and below the initial frequency. On a spectrum, this would appear as a doublet (two equal height peaks with the centre of the peaks equalling the chemical shift). For the example above when observing H_1 the splitting caused by H_2 will match the splitting caused by H_1 when observing H_2, allowing these two protons to be identified as close to one another. This coupling is not limited to two nuclei; if more magnetic nuclei are within range further splitting will occur. For organic molecules doublets, triplets, quartets and double doublets are most common. However, multiplets including sextets and septets are possible. Coupling of nuclei is an important phenomenon as it can be used to correlate spins in two-dimensional experiments. For this purpose, also the heteronuclear coupling *e.g.* between proton and carbon can be exploited using heteronuclear experiments.

1.2.2 Uses of NMR as a high-resolution technique for analysing HS

1.2.2.1 1D NMR

Through the versatility of different NMR spectroscopy methods (solid and solution-state NMR as well as gel-phase NMR) different types of HS can be explored, including DOM, FA and HA.[12] Solution-state NMR is the more common in terms of availability and has been used to gain a general view of the nature of structures within the sample as a whole. As shown (Figure 1.4), the ^{1}H spectra of Suwannee River dissolved organic matter can be acquired in different solvents to inspect what chemical moieties are present. Aliphatic, material derived from linear terpenoids (MDLT), carboxylic-rich alicyclic molecules (CRAM), carbohydrates, aromatic, amide and carbonyl moieties were all identified within the sample.[33]

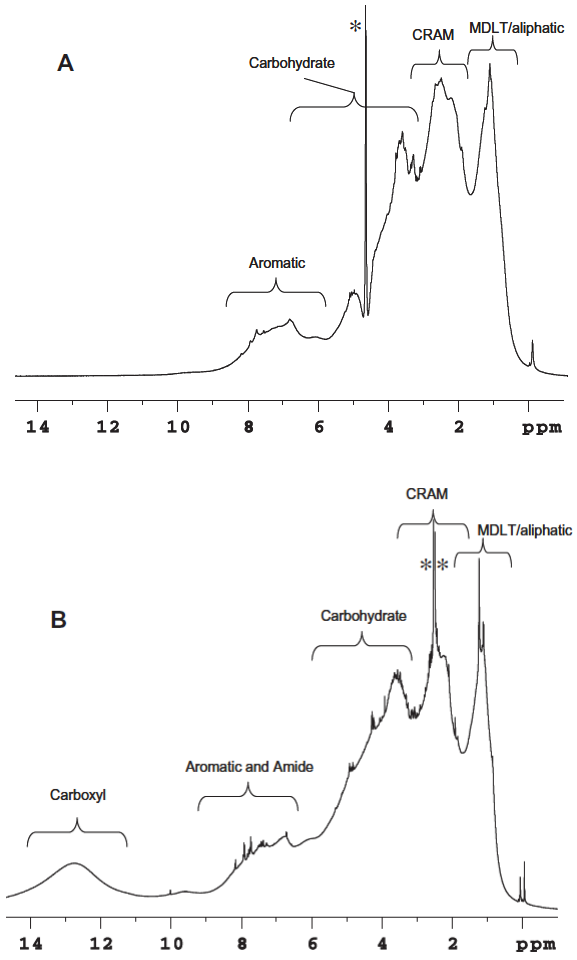


Figure 1.4 ^{1}H spectra of Suwannee river dissolved organic matter in A: D_2O/NaOD and B: DMSO-d_6.[33] Reprinted from Progress in Nuclear Magnetic Resonance Spectroscopy, 58/3-4, Simpson, A.J., McNally, D.J., Simpson, M.J., NMR spectroscopy in environmental research: From molecular interactions to global processes, 97-175, Copyright (2011), with permission from Elsevier.

Focusing on solution state NMR, for 1D analysis there are two main nuclei explored, proton and carbon. The spin half nuclei of carbon is ^{13}C, which has a 1.1% natural abundance, thus carbon NMR suffers from decreased sensitivity. For the study of HS by 1D NMR the resolution is not sufficient to separate the peaks from individual

compounds, so regions of the spectra are assigned to their related functionality and treated as a single entity. Figure 1.4 shows examples of these regions for ^{1}H spectra. The regions and associated chemistry for ^{1}H are given in Table 1.1 and for ^{13}C NMR are given in Table 1.2.[34]

Table 1.1 Assignment regions for ^{1}H NMR as suggested by Hertkorn.[34]

Range (ppm)	Assignment
10.0 – 6.0	$C_{aromatic}H$
6.0 – 4.7	acetal
4.6 – 3.1	CH_nO
3.1 – 1.95	C_f-CH_n
1.95 – 0.5	$C_{aliphatic}$-CH_n

Table 1.2 Assignment regions for ^{13}C NMR as suggested by Hertkorn.

Range (ppm)	Assignment
220 – 187	C=O
187 – 167	(C=O)-X-
167 – 145	$C_{aromatic}O/N$
145 – 108	$C_{aromatic}C/H$
108 – 90	O-C-O
90 - 47	$O(N_f)$-CH_n
47 - 0	CH_n

Hertkorn *et al.* in 2002 used these ranges to compare the overall sample chemistry of a HA and FA from the same peat source. The authors were able to identify that their peat HA contained a significant amount of basic aliphatic protons (24.9 % vs 14.2 %) whereas the peat FA contained more carbohydrates and benzene derived compounds

with little substitution (35.6 % *vs* 45.0 % and 15.6 % *vs* 18.2 %, respectively).[34] Similarly, the ^{13}C NMR spectra were analysed using the range given in Table 1.2. A larger difference in overall composition was seen than for the ^{1}H spectrum. Some notable observations were that the HA spectrum contained a higher amount of purely aliphatic, methyl esters and aliphatic methyl esters in comparison with FA. Improved resolution in the aromatic region of the spectrum was also seen for FA over HA.[34]

Further information can be gained from ^{13}C NMR by using DEPT editing. This allows for the separation of different carbon environments; primary (CH_3); secondary (CH_2); tertiary (CH); and quaternary (C-quat). An example of a ^{13}C NMR spectrum along with its DEPT spectrum for each environment is shown in Figure 1.5. Ivanona and Randall use this separation to quantify and compare the carbon environments, for the ranges given in Table 1.2, of FA and HA extracted from a High-field Grass soil.[35]

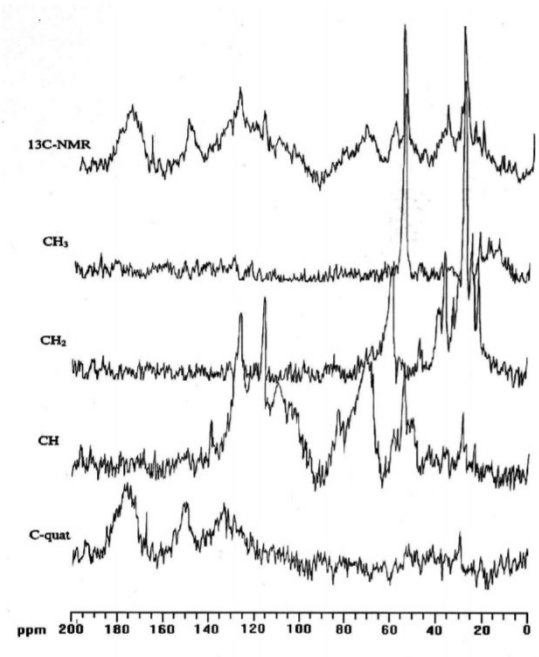


Figure 1.5 ^{13}C NMR and DEPT spectra showing separation of CH_3, CH_2, CH and C-quaternary carbon atoms (top to bottom) for HA extracted from High-field Grass soil.[35]

The benefit of using NMR as presented here is that the integrals of the peaks are quantitative in ^{1}H NMR and can be made, under appropriate experimental conditions quantitative for ^{13}C. Whilst the sensitivity and resolution of 1D NMR is insufficient to perform a thorough analysis of a single sample, by comparing samples it allows for broad comparisons to be drawn about the samples relative to one another. However, in order to gain a greater amount of information than offered by 1D NMR, higher dimensionality NMR experiments are required.

1.2.2.2 Multi-dimensional NMR spectroscopy

Progressing on from the information gained from 1D spectra, multi-dimensional NMR has been applied as a way to further 'separate' the HS compounds. In its most basic form, 2D ^{1}H-^{13}C HSQC spectra can be used to spread the signals along a second axis and partially separate them out. Ultimately this means that different hydrogen/carbon environments that are distinct can be separated. For example, signals from hydrogens that have the same chemical shift but are directly bound to carbon in differing environments can be separated. Figure 1.6 shows a 2D ^{1}H-^{13}C HSQC spectrum for Pacific Ocean DOM. The signals can be separated into seven partially overlapping regions: 1: methyl groups bound to carbon (solid line) and sulfur (hashed line); 2: methylene and methine far from any heteroatoms; 3: low intensity methoxyl; 4: α-CH_2 from proteins; 5: carbohydrate methylene systems; 6: carbohydrate methine systems and 7: anomeric moieties of carbohydrate monomers.[12]

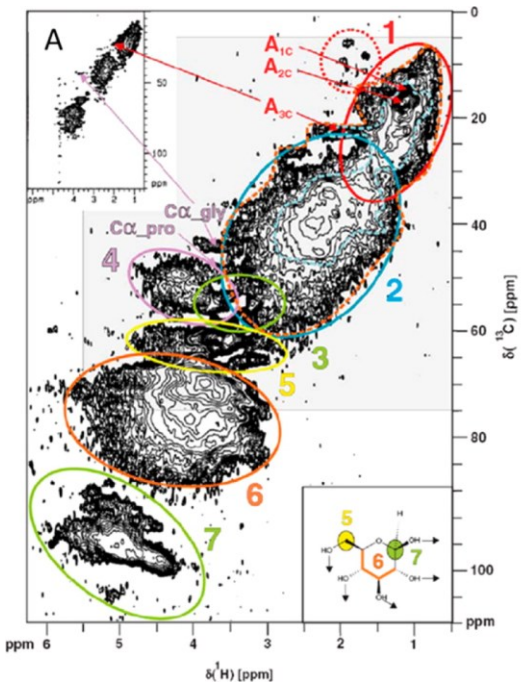


Figure 1.6 2D ^{1}H-^{13}C HSQC spectrum of Pacific Ocean DOM, see text for labelled regions.[12] Reprinted (adapted) with permission from Environ. Sci. Technol., 46/21, Simpson, A.J., Simpson, M.J., Soong, R., Nuclear Magnetic Resonance Spectroscopy and Its Key Role in Environmental Research, 11488-11496. Copyright (2012) American Chemical Society.

Other forms of 2D NMR used for HS analysis are correlated spectroscopy (COSY) and total correlation spectroscopy (TOCSY). Both experiments use the proton-proton coupling, to correlate spins, COSY initiates only the couplings between the directly coupled spins, while TOCSY can trace the spin systems of sequentially coupled protons. With the focus of these experiments being protons, they allow for a tentative analysis of the type of chemical species within the sample as opposed to just the functionality, when compared to HSQC. For example, Figure 1.7 shows COSY and TOCSY spectra for a peat FA and HA sample. Although the resolution is still not high

enough to resolve individual compounds, more specific regions can be identified. Examples of these regions from the COSY spectra are: carbohydrates without anomeric protons (C_C), aliphatics with one heteroatom functionality (D_C), anomeric protons of carbohydrate (E_C). For the TOCSY spectra, regions can be seen for: aliphatic chains (A_T), the Hα-Hβ of peptides (D_T), and aromatic protons that are ortho to one another (G_T).[34]

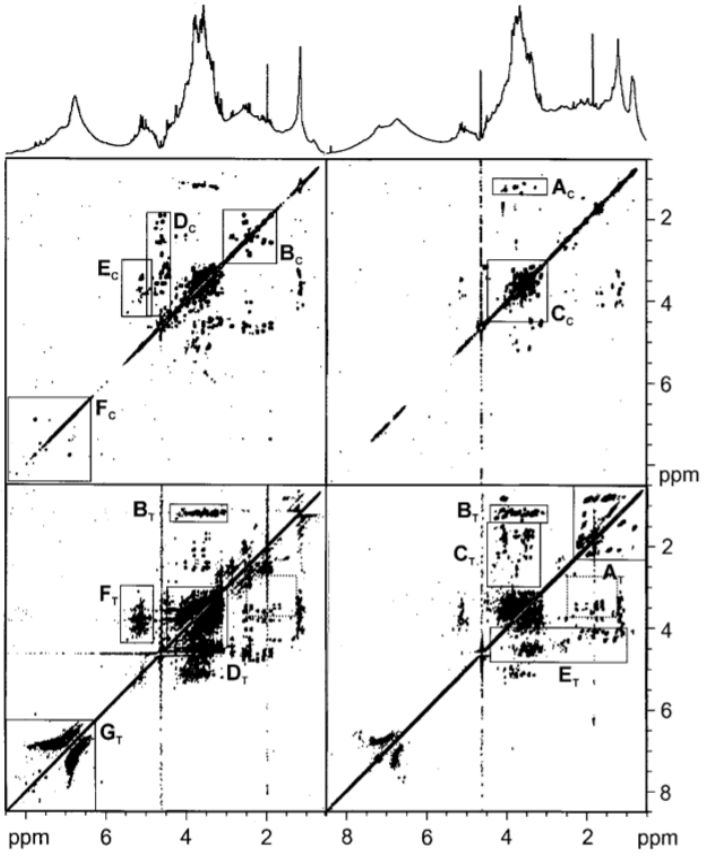


Figure 1.7 COSY (top) and TOCSY (bottom) spectra of peat FA (left) and peat HA (right).[34]

Although a greater level of separation is achieved by 2D NMR, which allows for a more sophisticated sample analysis, its most suitable use still is the comparison of

multiple samples. The requirement for the ability to study a single sample led deployment of even higher dimensionality approaches. In 2003, Simpson *et al.* demonstrated the use of 3D NMR on HS.[36] The authors used 3D HMQC-TOCSY experiments to separate the HS signals over an additional 3rd dimension, in this case two proton and one carbon dimensions (see Figure 1.8). By taking slices/planes of the carbon (F1) and proton (F2) dimensions a 2D spectrum with a reduced number of signals is obtained, where most of the complexity of the data is spread across the 3rd dimension. The 2D planes can then be used to associate coupled signal from the same molecule/molecular fragment. Figure 1.9 shows a plane taken at 1.3 ppm in the F3 dimension of Figure 1.8, this plane is used to identify molecular fragments containing aliphatic CH_2 groups.

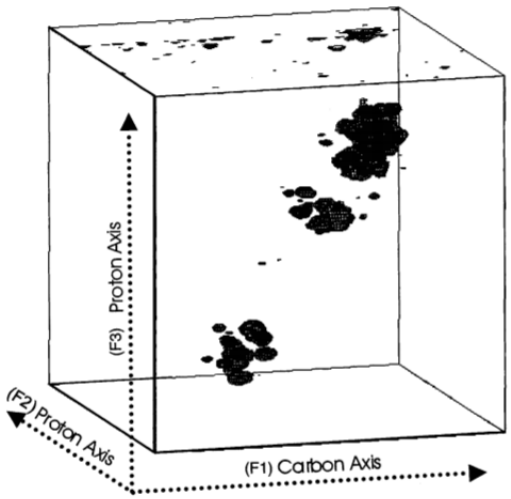


Figure 1.8 3D HMQC-TOCSY spectrum of pine forest FA.[36] Reprinted with permission from Environ. Sci. Technol., 37/2, Simpson, A.J., Kingery, W.L., Hatcher P.G., The Identification of Plant Derived Structures in Humic Materials Using Three-Dimensional NMR Spectroscopy, 337-342. Copyright (2003) American Chemical Society.

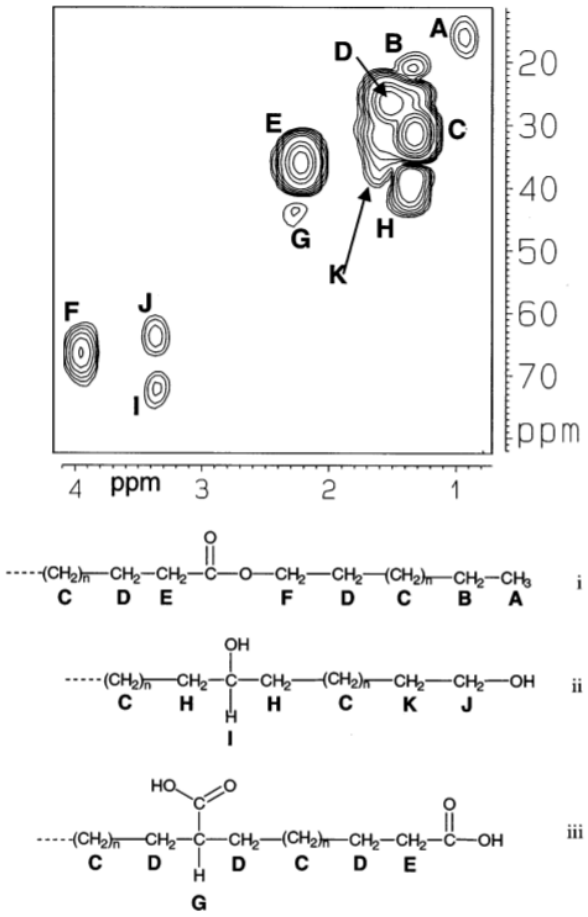


Figure 1.9 2D slice taken at 1.3 ppm in the F3 dimension of the 3D HMQC-TOCSY spectrum in Figure 1.10. Peaks are labelled and assigned to structural fragments i – iii.[36] Reprinted with permission from Environ. Sci. Technol., 37/2, Simpson, A.J., Kingery, W.L., Hatcher P.G., The Identification of Plant Derived Structures in Humic Materials Using Three-Dimensional NMR Spectroscopy, 337-342. Copyright (2003) American Chemical Society.

The molecular fragments produced shown in Figure 1.9 represent a big step towards single sample analysis. However, they all contain multiple $(CH_2)_n$ units, meaning what is presented is interesting functionality on a aliphatic backbone, rather than a molecular

fragment. Aromatic systems are more easily interpreted by NMR, this is because the chemical shift of aromatic protons can be more accurately predicted based on the functionality present on the ring, and the same is true in reverse, ring functionality can be reliably predicted based upon the aromatic protons chemical shift. A second plane taken at 6.9 ppm of the F3 dimension of Figure 1.8 was taken and two aromatic molecular fragments were identified, as shown in Figure 1.10.[36]

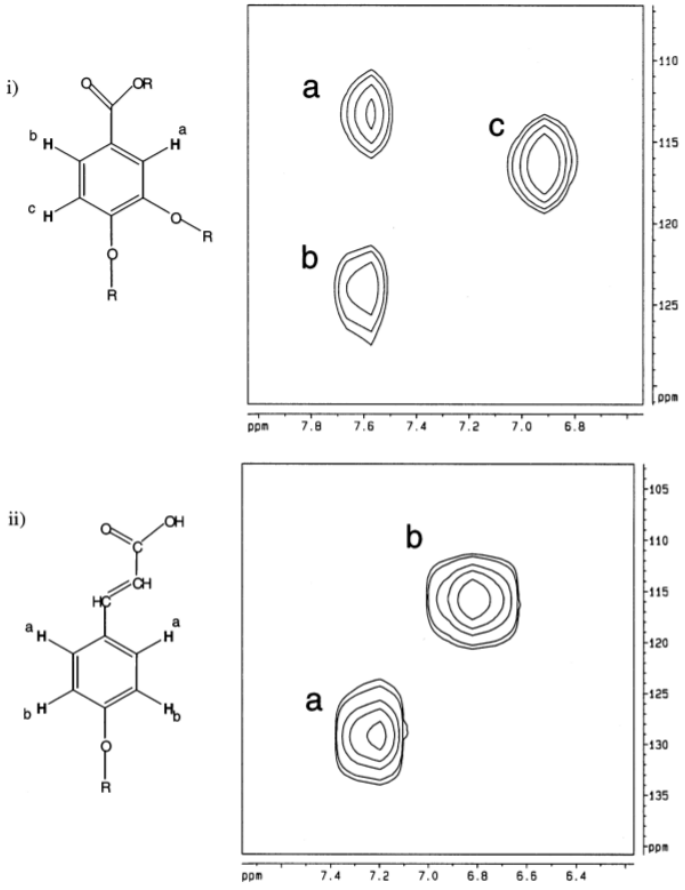


Figure 1.10 2D slice taken at 6.9 ppm in the F3 dimension of the 3D HMQC-TOCSY spectrum in Figure 1.10. Reprinted with permission from Environ. Sci. Technol., 37/2, Simpson, A.J., Kingery, W.L., Hatcher P.G., The Identification of Plant Derived Structures in Humic Materials Using Three-Dimensional NMR Spectroscopy, 337-342. Copyright (2003) American Chemical Society.

Higher dimensional NMR experiments above 3D are also possible, Bell *et al*. have performed 4D NMR experiments on a FA sample.[37] Although 4D NMR is harder to visualise graphically, practically it is similar to 3D NMR, the difference being that an additional dimension is used to spread the signal utilising the chemical shift of a forth nucleus. The annotated structure in Figure 1.11a shows an aromatic system with all the nuclei that can be sampled using the coupling types shown (starting from the $^{13}CH_3$ groups). Figure 1.11b is a 2D plane extracted from the 4D data; it has been greatly simplified comparted to the original 2D spectrum due to the addition of two dimensions of separation. The reduced spectrum, combined with the predictability of the chemical shift of signals on an aromatic ring allows for a small number of complete aromatic structures to be discerned. Whilst this 4D approach only identified a small number of simple compounds, it represents a giant leap in this field, as it showcases the first complete structures of HS identified by NMR.

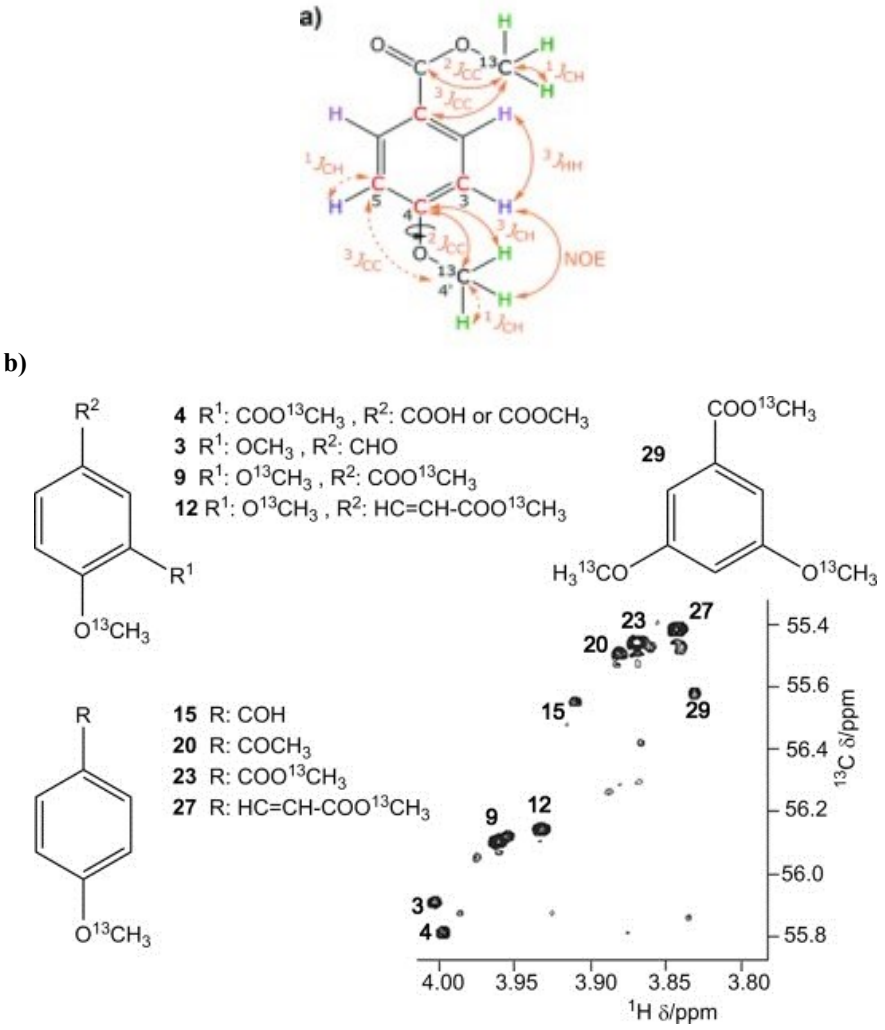


Figure 1.11 Example structure with couplings identifiable by 4D NMR, starting from the $^{13}CH_3$ group (a). A 2D ^{1}H, ^{13}C-HSQC spectrum extracted from a 4D data set alongside the associated structures that were identified (b).[37]

The references presented here show the current situation for NMR studies of HS. NMR lacks the resolution to identify single species and structures, and thus 1D and 2D NMR has most effectively been used in comparative studies between HS samples. Higher dimensionality, 3D NMR has been shown to increase separation for analysis of single

samples, however, thus far it has only yielded molecular fragments. Combinations of 3D and 4D NMR have yielded complete structures, however, only of a small number of simple compounds. In order to address insufficient resolution of NMR I explore the use of FTICR MS, the spectrometer at the forefront of high-resolution analysis, for analysis of HS.

1.3 Fourier transform ion-cyclotron resonance mass spectrometry

1.3.1 Principles of FTICR MS

FTICR MS is an analytical method first presented in 1974.[38] It provides information on the molecular mass of a compound by ionising compounds from solution before trapping them within a mass analyser. The mass analyser traps all ions within a magnetic field, of strength B. The ions travel in an arc with a radius of r (Equation 3). Therefore, if B is large enough, the resulting r will be small enough to cause the ion with a charge ze to become trapped in a circular motion with velocity (v_e) where the magnetic force, or Lorentz force (zev_eB) is balanced by the centrifugal force (mv_e^2/r) (Figure 1.12).

$$zev_eB = \frac{mv_e^{\,2}}{r} \tag{3}$$

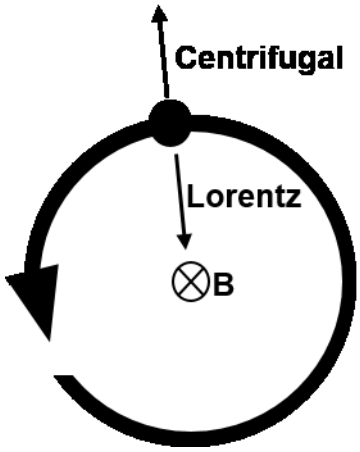


Figure 1.12 Circular orbit produced by the magnetic field (**B**) with the Lorentz and centrifugal forces opposing each another.

The cyclotron frequency (ω) is the frequency in Hz (ν) an ion completes a 2πr trajectory and so is independent of the velocity (v_e) because it only affects the cyclotron radius (r) of the ion (Equation 4). The frequency is inversely proportional to the *m/z* of the ion and can therefore be used to determine the mass of an ion.

$$\omega = 2\pi\nu = \frac{v_e}{2\pi r} \tag{4}$$

Once inside the mass analyser the ions are contained within the ICR cell between two perpendicular trapping electrodes and circulated around the direction of the strong magnetic field based on their cyclotron frequency. This cyclotron motion is the principal motion of the ions within the cell, however as the ions are only contained in the x and y planes (relative to the magnetic field) movement in the z direction is still possible (Figure 1.12). Therefore, a DC current is applied to both trapping electrodes

in order to contain the ions within a potential well, where they oscillate in a simple harmonic motion back and forth along the z axis.

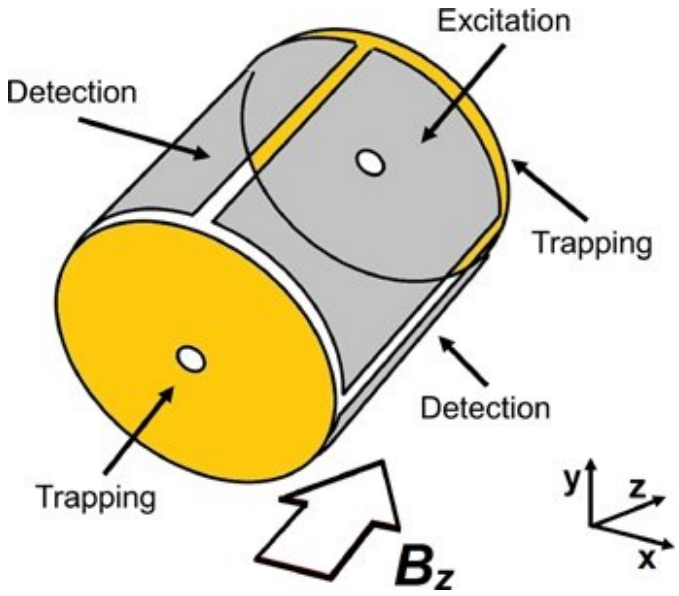


Figure 1.13 Schematic of an ICR cell.[39] Reprinted with permission from Mass Spectrometry Reviews, 33/5, Qi. Y. and O'Connor. P., Data processing in Fourier transform ion cyclotron resonance mass spectrometry, 333-352. Copyright (2014) John Wiley and Sons.

There is a third type of motion present within the cell, magnetron motion, which forms as a result of the combination of magnetic and electric fields upon the ions. The magnetic field constrains the ions in the xy-plane whereas the electric field constrains along the z-axis. However the electric field is not constant across the whole cell with a minimum at the centre and maxima at the trapping plates. The result is that the electric potential not only traps along the z-axis but also repels along the xy-plane, this

repulsion reaches a maximum at the centre of the cell, forcing ions closer to the centre of the analyser away. The magnetic field prevents ions being ejected into the analyser cell walls and the combination of these two forces is what produces magnetron motion. Magnetron motion is the precession of ions around the centre point of the cyclotron motion as they move around the centre of the cell. A combination of these two motions is shown in Figure 1.13 however the scale and frequency of the magnetron motion relative to the cyclotron motion has been greatly exaggerated for clarity.[40]

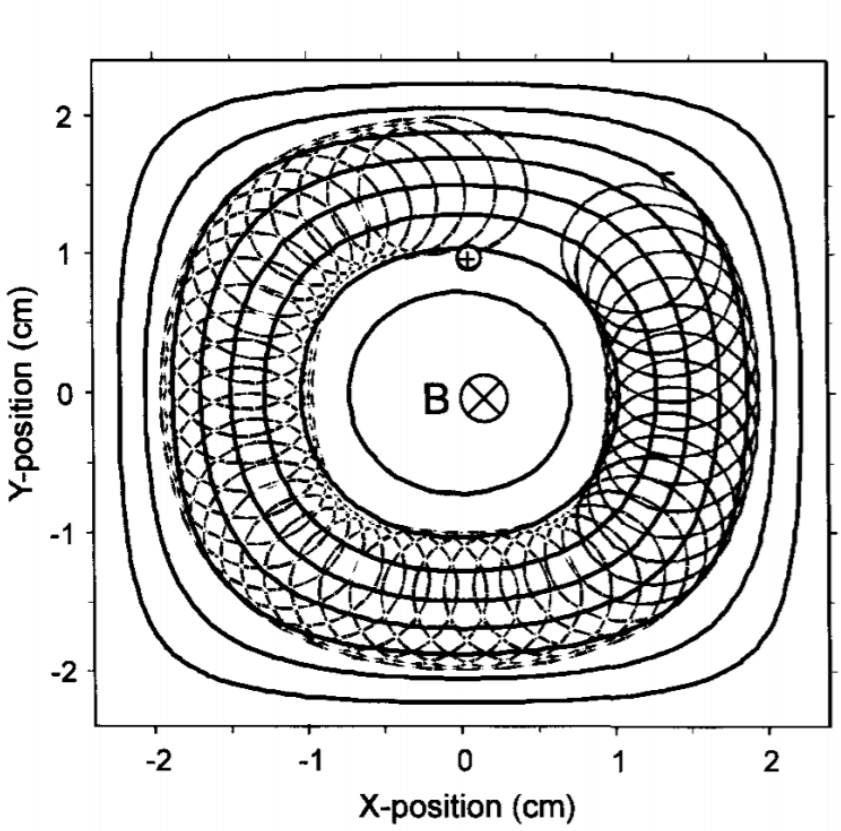


Figure 1.13 Ion trajectory showing cyclotron and exaggerated magnetron motion within a cubic cell.[41] Reprinted with permission from Wiley Books, Wilff. J.J., and Amster. I.J., Fourier Transform Ion Cyclotron Resonance and Magnetic Sector Analyzers for ESI and MALDI. Copyright (2012) John Wiley and Sons.

The ions trapped within the cell are typically at a lower kinetic energy (<1 eV), meaning they have small cyclotron radii. In order to detect the ions the radius of their orbit is increased by application of an RF frequency pulse across the excitation electrodes (Figure 1.13). This excitation increases the velocity of the ions that have a cyclotron frequency matching the pulse frequency and thus increases the radius of the ions trajectory whilst maintaining the cyclotron frequency (Figure 1.14A). It also causes all ions of the same *m/z* ratio to be brought into phase. The ions are accelerated to roughly 75% of the analyser radius for detection. After the RF pulse stops the ions continue to orbit around the magnetic field until collisions eventually cause all excitation energy to be lost. As this packet of spatially coherent ions passes by the detection electrode (which are connected to a ground) electrons are drawn towards that electrode and away from the other, this produces an AC current as the electrons move back and forth due to the cyclotron motion of the ions. This AC current change (or image current) is measured between the two electrodes, with the frequency of the current change from one electrode to the other and back again being equal to the cyclotron frequency (Figure 1.14B).

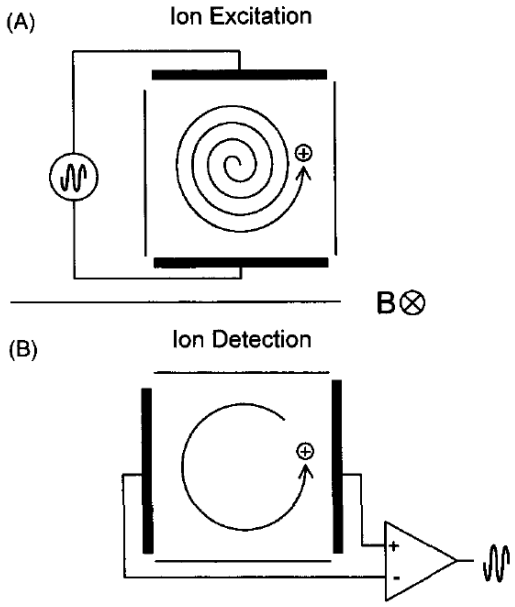


Figure 1.14 A) Ion trapped within analyser with cyclotron radius increasing due to matching frequency excitation pulse. B) Detection producing an image current from the cyclotron frequency of the excited ions.[41] Reprinted with permission from Wiley Books, Wilff. J.J., and Amster. I.J., Fourier Transform Ion Cyclotron Resonance and Magnetic Sector Analyzers for ESI and MALDI. Copyright (2012) John Wiley and Sons.

The ions are detected as they pass by each detector electrode without contacting the surface so the ions remain within the trap as they lose their excitation. This means no ion destruction should occur during the detection, allowing for signal detection over longer transients (2 – 3 s). For work with more than one mass it is required to excite all necessary *m/z* values near simultaneously, so rather than using an RF pulse at a set frequency a broadband frequency sweep ('RF chirp') is utilised. The image current for all *m/z* frequencies is simultaneously collected, with the final output from the analyser being a time domain signal for all *m/z* values swept over, this is referred to as the transient. The transient is sum of all the sinusoidal frequencies for every ion within the

analyser, a Fourier transformation converts this transient into the individual frequency components which can then be converted into the final mass domain spectrum.

1.3.2 Common ionisation methods in FTICR MS

1.3.2.1 Electrospray ionisation (ESI)

ESI involves spraying a solution out of a charged nozzle, due to the solution becoming charged a Taylor cone is formed that propels the solution droplets forward in a narrow stream.[42] This charged droplet undergoes desolvation up to the Rayleigh limit, fission of these droplets occurs to give single ion droplets. Further solvent evaporation produces charged analyte cations (positive mode) or anions (negative mode).[43] This method of ionisation is considered a soft method, meaning it doesn't cause extensive fragmentation of the molecule. This means that the ion under investigation is the ionised form of the original molecule, the molecular ion. In negative mode the ion is formed through loss of a single (or multiple) labile hydrogen(s). Positive mode involves formation of adducts such as $[M+H]^+$ or $[M+Na]^+$.

There are two possible mechanistic routes theorised for the formation of the ions, the charge residue model (CRM) and the ion evaporation model (IEM). The CRM involves the charges on the solvent droplet transferring to the analyte molecule within the droplet as the solvent evaporates (Figure 1.15a). The CRM mechanism for ionisation is used to explain how large macromolecular molecules gain multiple charges by ESI. This occurs when one molecule is held in a small amount of solvent that contains an excessive density of surface charge.[44] In the IEM, a charged analyte ion, which is produced within the droplet, is ejected directly from the surface of the droplet. This ejection is considered the evaporation of a small ion. The IEM is used to

describe how ions of small molecules are produced by ESI (Figure 1.15b).[45] Whether an ion is produced by the CRM or IEM depends on its, however, the mass at which an ion is no longer ionised by IEM, but instead by CRM, is unknown. There are likely molecules that produce ions via both methods.

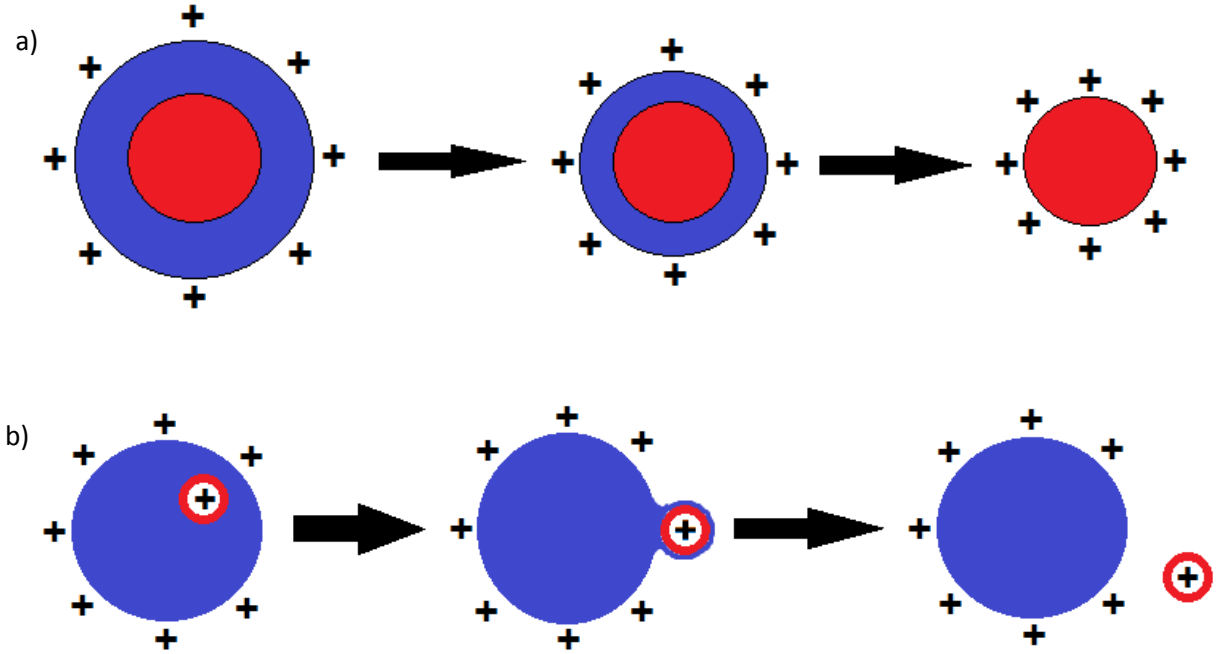


Figure 1.15 Pictorial representation of: a) CRM producing an M^{8+} ion and b) IEM producing a singly charged ion. Solvent – blue, ionised molecule – red.

1.3.2.2 Matrix-assisted laser desorption/ionisation (MALDI)

MALDI is another soft ionisation method that stems from a need to ionise large molecules that often suffer from poor solubility, such as proteins.[46] It works by imbedding the molecule in a large excess of matrix. This matrix/sample mixture is dried onto a MALDI plate and mounted into the instrument. Once mounted the target spots are irradiated using laser light, typically in the UV region, which causes the

sublimation and ionisation of both matrix and sample molecules in a single step. The mechanism for the formation of ions in MALDI is not fully understood.[47-49] It is a common belief that proton transfer occurs between the sample and matrix during either the solid or gas phase, however, this is yet to be proven.[50] The process of sublimation is better understood and is attributed to two mechanisms, "desorption" and "ablation". The properties of the laser decides whether material is desorbed or ablated from the MALDI plate. Figure 1.16 shows simulations produced for these two mechanisms; desorption causes a steady release of material from the outer surface of the sample, containing no small or large clusters; ablation is caused by overheating the core of the sample by the laser, leading to nucleation, which forces material from the surface containing clusters, particles and single molecules.[48, 51]

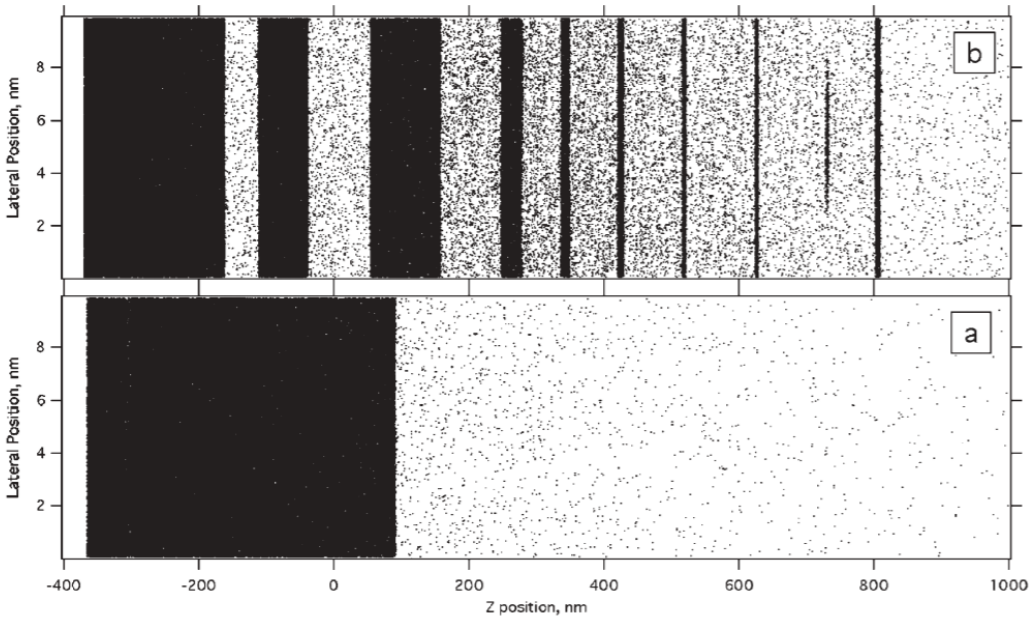


Figure 1.16 A single frame from a simulation of MALDI following desorption (a) and ablation (b) mechanisms[51] Reprinted with permission from Analyst, 131/9, Knochenmuss. R., Ion formation mechanisms in UV-MALDI, 966-986. Copyright (2006) Royal Society of Chemistry.

Laser Desorption/Ionisation (LDI) is a matrix-free variation of MALDI. Laser pulses are fired directly onto the sample, producing a microplasma at the surface of the sample where reactions between ions and neutral molecules occur, producing gas phase ions. It is an efficient way to produce ions and can be used to study the surface of materials. However, increased laser power[52] required to ionize analytes compared to MALDI[53] it is believed to cause sample fragmentation of species above *m/z* 500.[50]

1.3.3 Data handling and visualisation

1.3.3.1 van Krevelen diagram

A commonly used method for visualising large datasets of HS MS are van Krevelen diagrams, introduced in 1950[58], which are plots of the H/C ratio vs the O/C ratio for all the formulae assigned. Regions can be identified and roughly labelled as to give an idea of the type of compounds present in the sample. These regions consist of lipids, proteins, aminosugar, carbohydrates, condensed hydrocarbons, lignin derived molecules and tannins, as shown in Figure 1.17.[59]

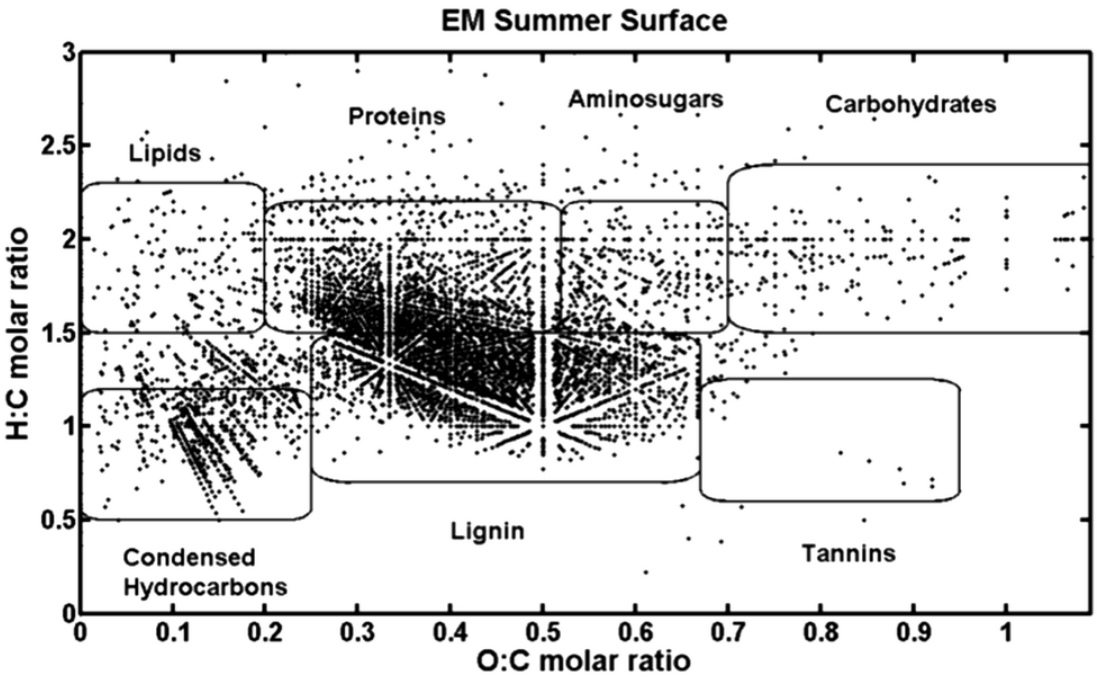


Figure 1.17 van Krevelen diagram of compounds identified from DOM extracted from Lake Superior.[59] Reprinted with permission from Environ. Sci.: Processes & Impacts, 16/9, Minor. E.C., Swenson, M.M., Mattson. B.M., Oyler. A.R., Structural characterization of dissolved organic matter: a review of current techniques for isolation and analysis, 2064-2079. Copyright (2014) Royal Society of Chemistry.

Using these rough outlined areas, a suggestion of the chemical nature of the sample can be concluded. Figure 1.18 shows a van Krevlen diagram for HA from an ancient, brown wood sample. The highlighted regions are described as A: charcoal based structures and aromatic systems due to the low hydrogen content, which are probably condensed structures. B: is typically what is expected to be seen from lignin and the direct derivatives of lignin. C: is likely to be cellulose and similar structured biomolecules.[60]

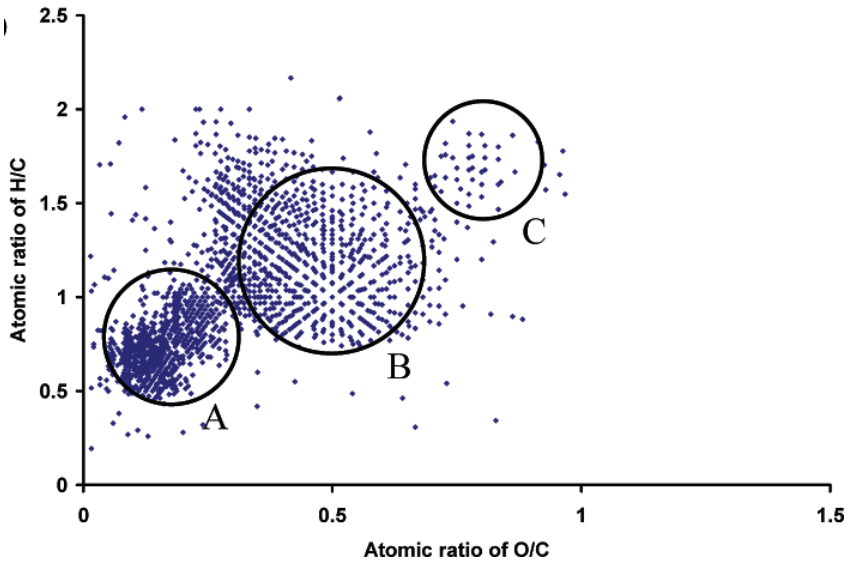


Figure 1.18 van Krevlen diagram of brown wood HA, description in text.[60] Reprint adapted with permission from Anal. Chem., 75/20, Kim. S., Kramer. R.W., Hatcher. P.G., Graphical Method for Analysis of Ultrahigh-Resolution Broadband Mass Spectra of Natural Organic Matter, the Van Krevelen Diagram, 5336-5344. Copyright (2003) American Chemical Society.

The weakness of using van Krevelen diagrams is a consequence of the associated limitations with using H/C ratios and O/C ratios as separation variables, that act as a data reduction agents. Non-identical formulae can have identical H/C and O/C ratios, for example, $C_6H_{12}O_6$, $C_9H_{18}O_9$ and $C_{12}H_{24}O_{12}$ all have a H/C ratio of 2 and an O/C ratio of 1. These separate formulae will occupy the exact same point on the van Krevelen diagram and are indistinguishable from one another as a result. Additionally, analysis based on the pre-defined regions of a van Krevelen diagram must be interpreted with caution. Figure 1.19 show two structural isomers for a lignin model structure and a CRAM model structure, despite being significantly structurally

different, they are structural isomers with the formula $C_{26}H_{32}O_{11}$.[61] Consequentially, these two structures will both be represented by the same point on a van Krevelen diagram despite their significant structural differences.

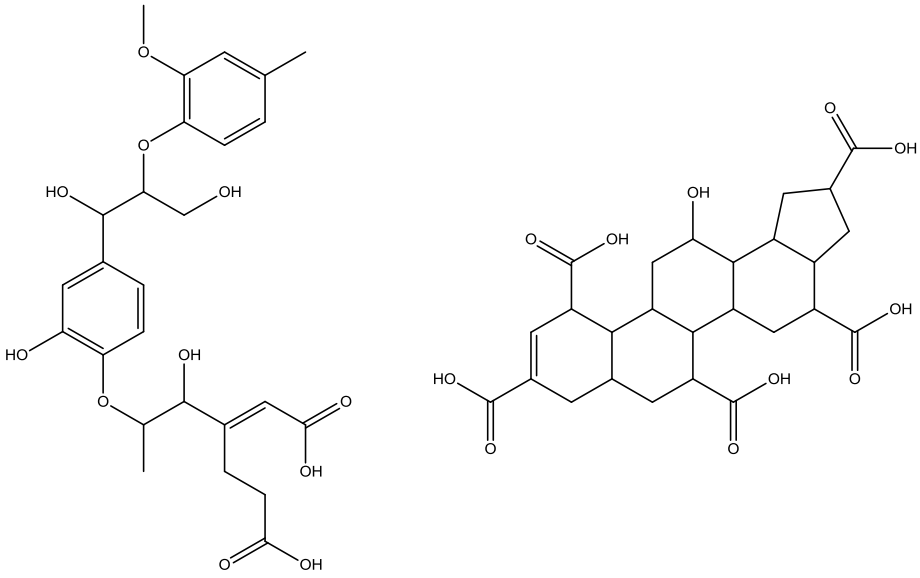


Figure 1.19 Two structure isomers for the formula $C_{26}H_{32}O_{11}$. Left shows the structure of a model lignin compound, and right the structure of a model CRAM compound.[61]

1.3.3.2 Double bond equivalence and aromaticity index

Double bond equivalents (DBE) is a mathematical metric for characterising a molecule that requires only the molecular formula to be known. It is calculated (for species containing only CHON) using the Equation 5. The final number from this calculation is a sum of the number of double bonds, rings and two times the number of triple bonds within the molecule. These properties give it merit for use with FTICR MS data of complex mixtures, where many formulae can be gathered. DBE can provide some level of structural interpretation, reporting on the unsaturation and aromaticity of the molecule. For example, benzene has a DBE of 4 whereas hexane has a DBE of 0.

$$DBE = C - \frac{H}{2} + \frac{N}{2} + 1 \quad (5)$$

Whilst DBE functions mathematically as intended, chemically it falls short of expectation, as it does not correctly describe oxygen functionality within a molecule. For example, the formula C_3H_6O has a DBE of 1, however, depending on whether the oxygen in the structure is π bound (double bonded) or σ bound (single bonded) decides whether the molecule contains a C=C as suggested by its DBE. Koch and Dittmar in 2006 acknowledge this issue and designed a metric to account for oxygen within a molecule and better predict aromaticity.[62] Aromaticity index (AI) was their solution, calculated as a ratio of the modified DBE (DBE_{AI}) and carbon number (C_{AI}) (Equation 8). DBE_{AI} is calculated similarly to DBE, however its variables have been adjusted to account for the number of oxygens, as well as to correctly describe the true effects of other heteroatoms such as sulfur (Equation 6). C_{AI} is the carbon number reduced by the potential number of double bonds to heteroatoms (Equation 7)

$$DBE_{AI} = 1 + C - O - S - 0.5H \quad (6)$$

$$C_{AI} = C - O - N - S - P \quad (7)$$

$$AI = \frac{DBE_{AI}}{C_{AI}} = \frac{1 + C - O - S - 0.5H}{C - O - S - N - P} \quad (8)$$

The resulting value from these calculations, AI, is either negative (and treated as if it were 0) or is a value between 0 and 1. Boundaries can be placed between these values to describe aromaticity: AI ≤ 0.5 is considered non aromatic; 0.5 < AI < 0.67 is aromatic; and 0.67 ≤ AI is a condensed aromatic. The issue with AI is that is accounts for all oxygen in the molecule being π bound. This isn't the case for some basic functionalities like carboxylic acid and ester groups, as both contain one π bound and

one σ bound oxygen. Additionally, there are as many π bound oxygen groups (ketones and aldehydes) as there are sigma bound (alcohol and ether). Therefore, especially for large molecules containing several oxygens, a safe approximation to make is that only half the total number of oxygens are π bound. The result of this adjustment is AI_{mod}, shown in Equation 9.[62]

$$AI_{mod} = \frac{1 + C - 0.5O - S - 0.5H}{C - 0.5O - S - N - P} \quad (9)$$

1.4 FTICR MS as a tool for analysing HS

1.4.1 Ionisation of HS in FTICR MS

The majority of MS studies of HS use ESI due to its ability to ionise a large number of polar compounds present in these complex mixtures.[63] However, alternative ionization techniques, including APPI, APCI and MALDI, have also been used in a few cases to provide a complementary picture of HS sample composition.[64-66] The most comprehensive comparison to date comparing ESI, APPI and APCI ionization of SRFA found that the choice of the ionization method had major impact on the nature of observed compounds. All methods were only able to ionize parts of the sample. Some molecules were common to all three ionization methods; some were shared between two, while some were unique to a specific ionization method.[64]

LDI is significantly less used in complex mixture analysis.[67, 68] LDI was thought to cause excessive fragmentation of HS.[66] However, this notion originated from early works on HS, which assumed that HS were composed of large polymeric molecules.[69]

When low (*m/z* 200 – 700) and higher (*m/z* 1000 – 1600) molecular weight peaks were observed in the LDI (but also ESI) MS spectra, the low mass peaks were initially interpreted as fragmentation products.[70] However, subsequent studies have shown that HS are aggregates of small molecules.[15]

1.4.2 FTICR MS for the analysis and comparison of HS samples

The use of FTICR MS to analyse HS has grown since 2002, when Stenson *et al.* used the technique to reinforce the challenge to the theory that HS consisted primarily large species (>*m/z* 2000), despite methods other than mass spectrometry, producing consistent results confirming that HS does contain large species.[71] The authors instead found that their HS sample (SRFA) contained many small molecules, and inferred that they must appear larger due to aggregation. The mass range of these molecules ranged from *m/z* 300 – 1000 with a maximum abundance at approximately *m/z* 600. Figure 1.20 shows the spectra collected using negative mode ESI along with expanded sub-spectra used to display characteristic trends.

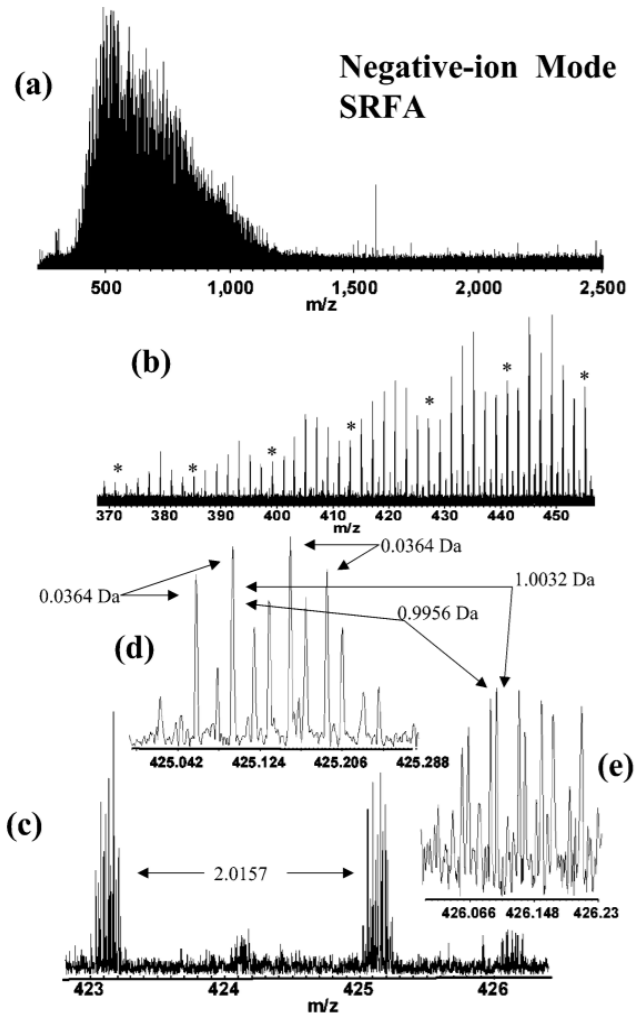


Figure 1.20 Negative mode ESI FTICR spectrum of SRFA at 5.5 mg/mL. (a) is the broadband spectrum and (b - e) show expanded regions.[71] Reprint adapted with permission from Anal. Chem., 74/17, Stenson. A.C., Landing. W.M., Marshall. A.G., and Cooper. W.T., Ionization and Fragmentation of Humic Substances in Electrospray Ionization Fourier Transform-Ion Cyclotron Resonance Mass Spectrometry, 4397-4409. Copyright (2002) American Chemical Society.

Thousands of peaks were identified for each spectrum and formulae were assigned to many of them. From this, major trends in formulae across the spectra were identified. These major trends (Figure 1.20) were described by repeating units of: 14.0156 *m/z* units, caused by an addition one of CH_2 group; 1.0034 *m/z* unit, which can be used to identify single ^{13}C isotope peaks; and 0.0364 *m/z* unit, which is a formulaic exchange of CH_4 vs O.[71]

Following the introduction of FTICR MS as a viable option for analysis of complex mixtures, research of HS using the technology began to increase.[72] The main metrics for data interpretation and visualisation of extensive formulae tables were van Krevelen diagrams (Section 1.3.4.1) and DBE plots (Section 1.3.4.2).[73, 74] However, similarly to NMR, the information gained from these analyses was quite broad and lacked the specificity required to accurately describe the sample. As a result, again similarly to NMR, the MS techniques were used to compare samples, as this proved to produce more interesting results. Figure 1.21 shows a van Krevelen diagrams for comparison of DOM from the Weddell Sea and from mangrove extracted pore water, this comparison allows for more specific differences to be identified. In this case, the Weddell Sea DOM clearly has more condensed character but less lipid and tannin character.[75]

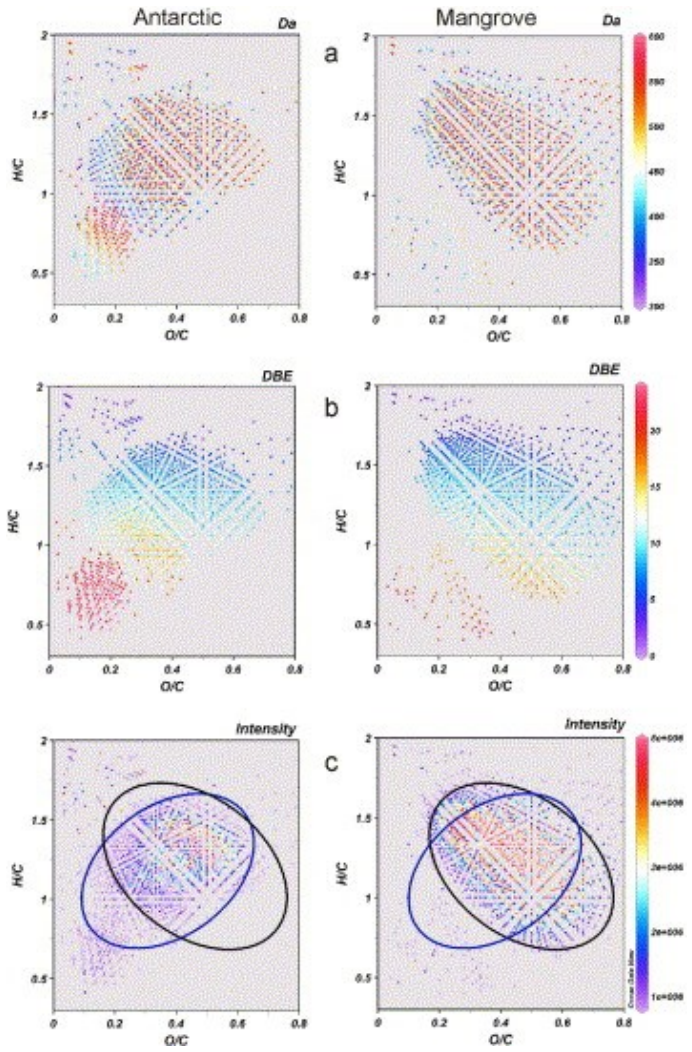


Figure 1.21 van Krevelen diagrams for ESI FTICR MS spectra of Weddell Sea (left) and Brazilian mangrove pore water (right). Z-axis fitted to molecular weight (a), DBE (b), and absolute peak intensity (c).[75] Reprint adapted with permission from Geochimica et Cosmochimica Acta, 69/13, Koch. B.P., Witt. M., Engbrodt. R., Dittmar. T., and Kattner. G., Molecular formulae of marine and terrigenous dissolved organic matter detected by electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry, 3299-3308. Copyright (2005) Elsevier.

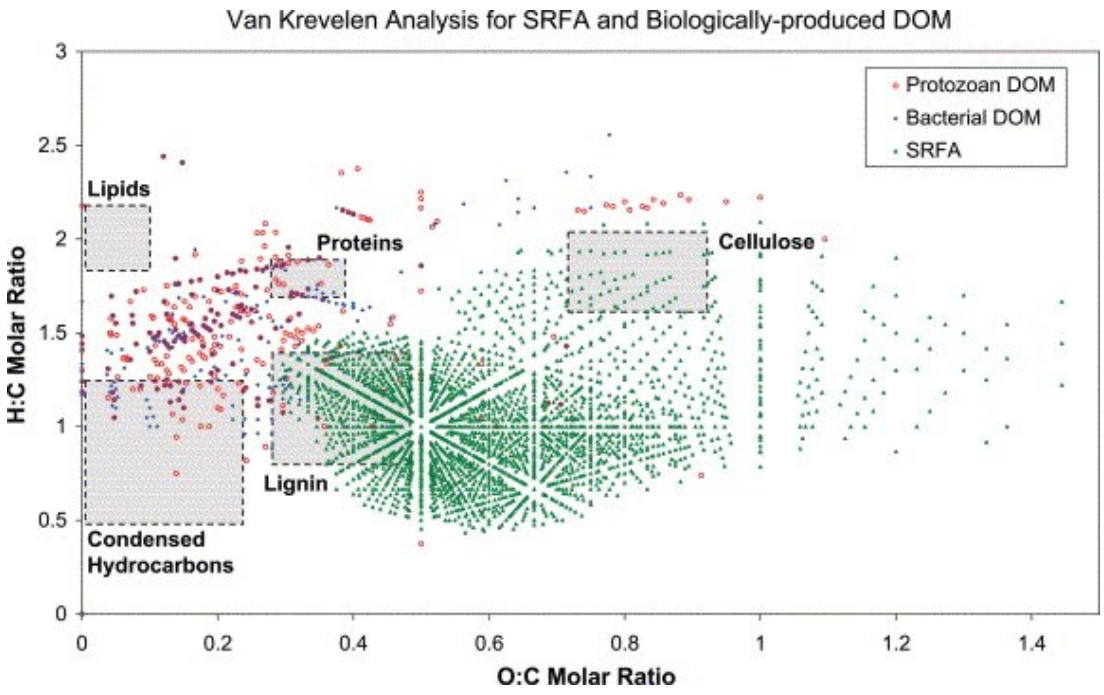


Figure 1.22 van Krevelen diagram for ESI FTICR MS of protozoan and bacterial modified DOM and SRFA. Highlighted regions shown general regions of compound classes.[76] Reprinted with permission from Marine Chemistry, 92/1-4, Kujawinski. E.B., Del Vecchio. R., Blough. N.V., Klein. G.C., and Marshall. A.G., Probing molecular-level transformations of dissolved organic matter: insights on photochemical and protozoan modification of DOM from electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry, 23-37. Copyright (2004) Elsevier.

The analysis technique shown here can also be performed without having two similar samples to compare. Figure 1.22 shows a van Krevelen diagram of two DOM samples, both of which have been modified. One by a protozoan culture and the other by a bacterial culture. These van Krevelen diagram are overlaid onto the van Krevelen diagram of SRFA. In this case SRFA is acting like a chemical standard for the DOM

degradation. The authors attempt to identify what kind of functionality and chemistry is produced by the degradation in comparison to standard molecules seen in HS. From Figure 1.22 it can clearly be seen the resultant modified DOM was significantly different to SRFA, being more aliphatic, less oxygenated and likely containing more nitrogen.[76]

Attempting to improve the information gained from this comparison, Kellerman *et al*. performed one of the most intensive and thorough pieces of research for this field to date. They extracted DOM from 109 Swedish lakes and analysed them using Excitation-emission matrix spectroscopy (EEM) and FTICR MS. PARAFAC, a multi variant analysis method used for chemometrics, was performed on the EEM data to model the 6 main variables of the fluorescence data. Spearman correlation was performed between these PARAFAC models and the FTICR MS formulae assignments. From this, the authors generated van Krevelen diagrams that relate the formulae identified to the fluorescence. Based on this they were able to identify molecular types (for example, vascular plant-derived polyphenols) and relate them to biogeochemical factors and processes observed in the lakes. The PARAFAC models and the Spearman correlated van Krevelen diagrams can be seen in Figure 1.23.[77]

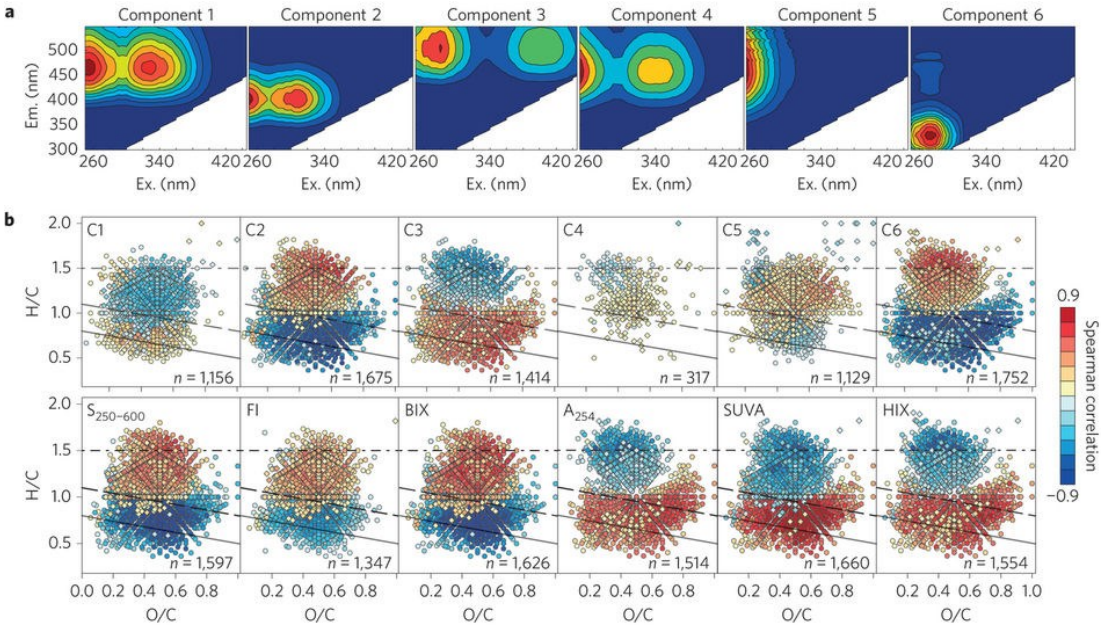


Figure 1.23 (a) Models of EEM data generated using PARAFAC, and (b) Spearman correlations between FTICR MS compounds and fluorescent compounds.[77] Reprinted by permission from: Springer Nature, Nature Geoscience, Persistence of dissolved organic matter in lakes related to its molecular characteristics, Kellerman. A.M., Kothawala. D.N., Dittmar.T., and Tranvik, L.J., Copyright (2015).

1.4.3 Towards structure determination of molecules by FTICR MS

The aforementioned works all rely on comparing samples; they all roughly hit a limit on what information can be gained based purely on chemical formulae. In order to increase the information content of FTICR mass spectra, new methodology must be developed. One such approach is molecular isolation and fragmentation within the instrument, this act. Witt *et al*. demonstrated this approach on negative mode ESI FTICR MS of SRFA. In order to isolate ions from such a complex mixture a quadrupole was initially used to reduce the ions to a 20 Da window, followed by

correlated harmonic excitation field[78] isolation to reduce the window to 1 Da. Long, sharp excitation pulses were then used to selectively excite a single mass from the 1 Da window. Figure 1.24 shows the *m/z* 365 region and the individual masses isolated using this method.[79]

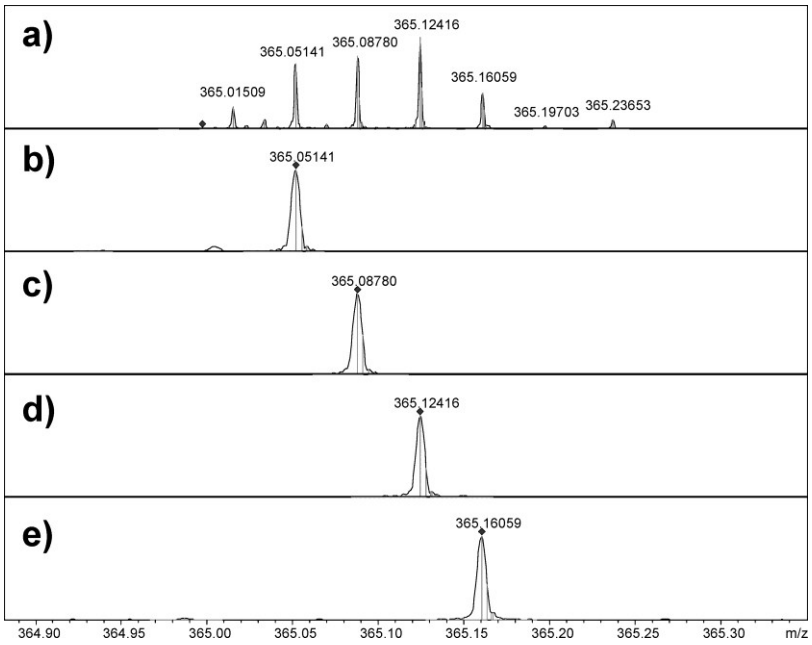


Figure 1.24 ESI FTICR MS of SRFA with the mass range at *m/z* 365 isolated (a). The 4 major signals from this range having been selectively excited (b – e).[79] Reprinted with permission from Anal. Chem., 81/7, Witt. M., Fuchser. J., and Koch. B.P., Fragmentation Studies of Fulvic Acids Using Collision Induced Dissociation Fourier Transform Ion Cyclotron Resonance Mass Spectrometry, 2688-2694. Copyright (2009) American Chemical Society.

Once a single mass peak has been selectively excited, it can be fragmented using in-cell collision induced dissociation (CID) and analysed based on its fragmentation pattern. Figure 1.25a shows the fragmentation of the peak isolated in Figure 1.24c.

From the fragmentation pattern the structure was predicted, as shown in Figure 1.25b.[79]

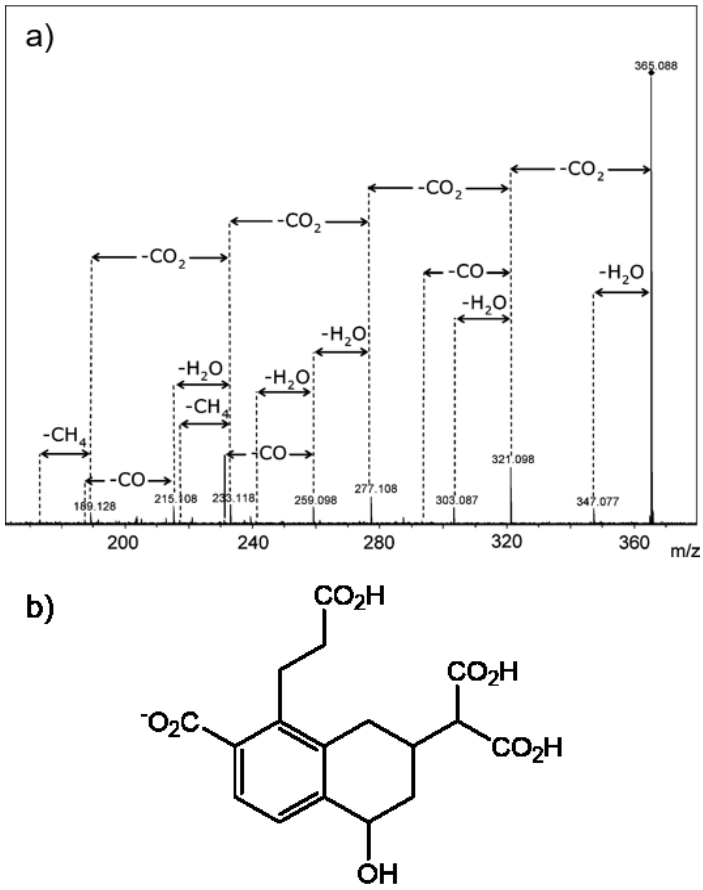


Figure 1.25 (a) In-cell CID spectrum of peak at *m/z* 365.088 (Figure 1.24c) with losses annotated. (b) predicted structure of the molecule isolated in (a) based upon fragmentation pattern.[79] Reprint adapted with permission from Anal. Chem., 81/7, Witt. M., Fuchser. J., and Koch. B.P., Fragmentation Studies of Fulvic Acids Using Collision Induced Dissociation Fourier Transform Ion Cyclotron Resonance Mass Spectrometry, 2688-2694. Copyright (2009) American Chemical Society.

The process of isolation and fragmentation of sample ions within the spectrometer can be described as 2D FTICR MS or FTICR MS/MS. In 1984, Marshall *et al*. discovered that excited ions could be "de-excited" using an inverse of the excitation pulse, and theorised that this form of selective in-cell isolation could have major advantages for FTICR MS/MS studies.[80] An experiment designed to perform in-cell isolation by using selective de-excitation was designed by 1993 and called stored-waveform ion modulation (SWIM).[81] Due to limited computer power at the time, the performance and application of using this method was low. However, since 2000, computing power has increased dramatically and the acquisition systems on instruments has become digitized, increasing their stability. The result is that FTICR MS/MS using selective de-excitation is now more reliable and controllable.[82] The stability of in-cell isolation is a consequence of gas-free fragmentation, as collisions with gas molecules can disrupt the ion packets, causing them to fall out of coherence. Infrared multiphoton dissociation (IRMPD) and electron-capture dissociation are both suitable gas-free alternatives to CID. Using SWIM to isolate all the precursor ions in a sample and then to fragment them using IMRPD, the fragments and the precursor ion can be detected and displayed in a 2D plot (Figure 1.26). The authors noted that the required computational power remains too high even for today's standard. Consequently the resolution of the data acquired is well below what would be expected for FTICR MS.[82]

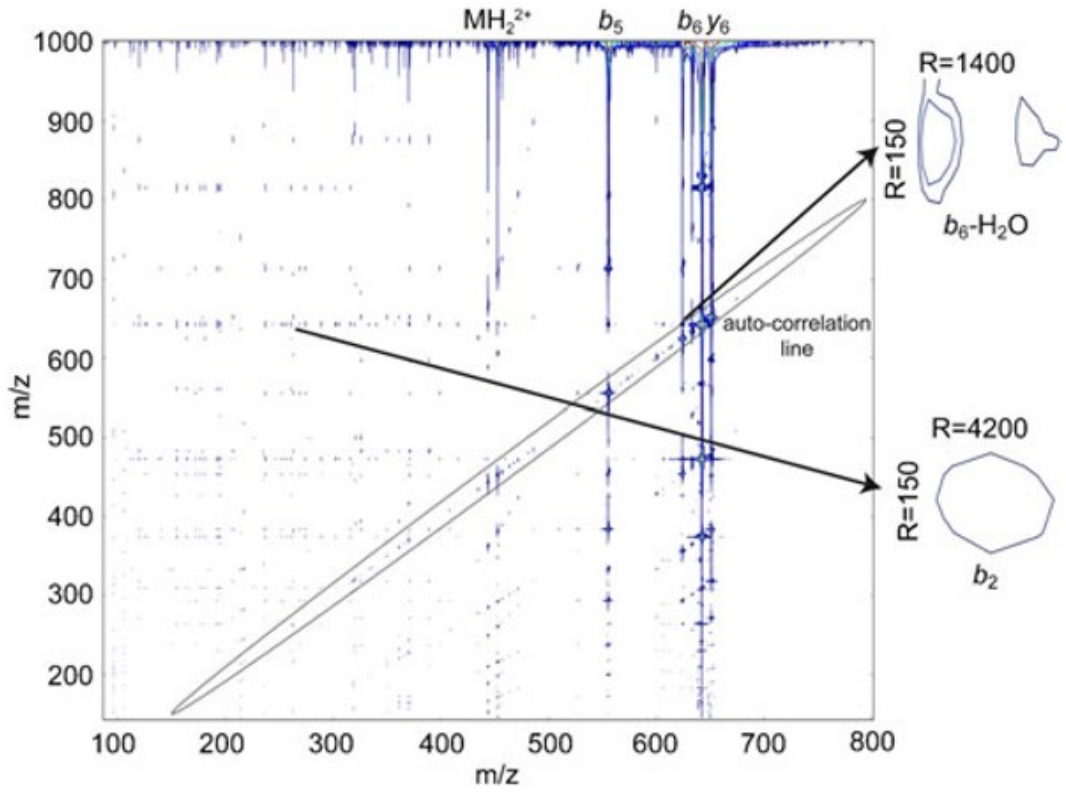


Figure 1.26 2D FTICR MS spectrum of bradykinin acquired using SWIM isolation and IRMPD fragmentation.[82] Reprinted by permission from: Springer Nature, Analytical and Bioanalytical Chemistry, Towards analytically useful two-dimensional Fourier transform ion cyclotron resonance mass spectrometry, van Agthoven. M.A., Delsuc. MA., Bodenhausen. G., Copyright (2012).

Whilst isolation and fragmentation in this manner has the potential to produce structures of molecules in HS, it doesn't fully overcome the complex nature of HS. The isolation step can successfully isolate a single mass peak; however, this is not necessarily a single compound. Fitting with the complex nature of HS this single mass peak is more likely to be a combination of multiple structural isomers. So, the isolation,

and then fragmentation steps, still occur on multiple compounds; resulting in the fragmentation pattern being a combination of multiple structures fragmentation.

1.5 Chemical modification

Prior to the understanding that HS were aggregates of small molecules, and the drive since then to fully elucidate the chemistry of these mixtures, oxygen based functionality was an area of key interest. Due to the natural propensity for complex formation and acidity of oxygen based groups they were of key interest to those investigating the biogeochemical properties of HS. One research direction that attempted to probe the oxygen functionality of HS was using chemical modification. Methylation reactions were used to modify carboxylic acid and alcohol groups within HS in order to label select functionalities. Some work focussed just on a single methylation reagent, methyl iodide (MeI).[83] MeI reacts with at all labile proton sites via a S_N2 substitution reaction and adds a CH_3 group. Other work used a combination of methylation agents, initially using diazomethane to selectively methylate carboxylic acid groups, then using MeI to methylate the remaining alcohol groups.[84, 85] Methylated products were then analysed by NMR, however, higher dimensionality NMR and the required instruments to perform them weren't available at the time so only simple 1D ^{13}C NMR was performed.[86, 87] Recent work by Bell *et al.*[88] has re-evaluated methylation using MeI by using ^{13}C enriched MeI to increase the methyl groups sensitivity for study by NMR spectroscopy. This enabled the authors to perform multi-dimensional NMR from a known starting point within the molecule. Figure 1.27 shows the nine compounds from their methylated model mixture, with the nuclei

identified using a 3D IPAP INEPT-INADEQUATE-HSQC spectrum presented in bold.

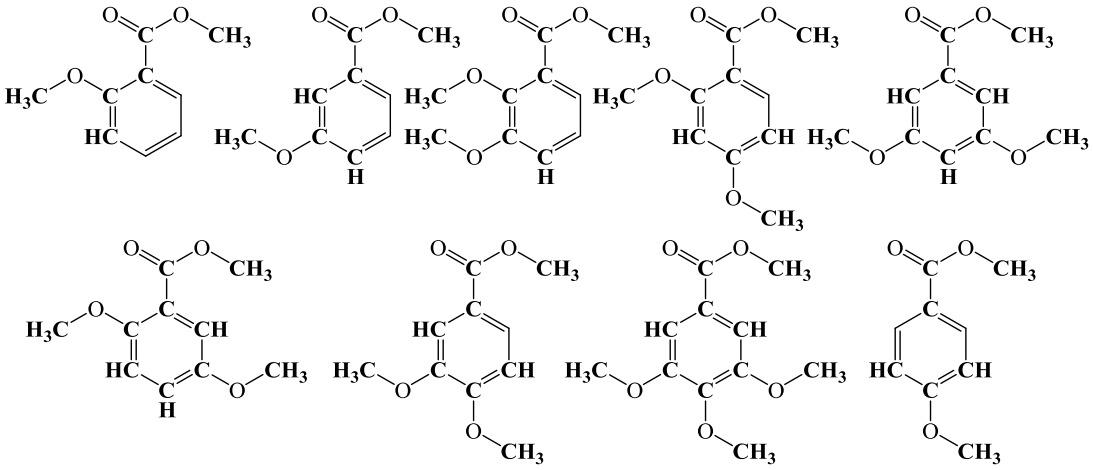


Figure 1.27 Model mixture compounds methylated using ^{13}C enriched MeI. Nuclei highlighted in bold identified via chemical shift from a 3D IPAP INEPT-INADEQUATE-HSQC spectrum.[88]

Standard methylation as previously discussed is not suitable for analysis of HS by FTICR MS, as the methyl labels just add carbon and hydrogen, similar to what is already present. In order for a chemical label to be viable for analysis by FTICR MS it must contain an alien atom. Kostyukevich *et al*. attempted such an approach by using deuterium.[89] The authors achieved deuterium labelling of HS by spraying the sample through a heated mist of D_2O within the source of the instrument. This method causes labile protons to exchange for deuterons as the sample is sprayed into the spectrometer, and is aimed at all exchangeable protons (OH, COOH, NH). The mass spectrum can then be analysed and formulae identified along with the number of deuterons it

contains, which acts as a count of the number of labile protons in the parent compound. Figure 1.28 shows a van Krevelen diagram of SRFA labelled in this manner alongside with the number of exchanges observed for each identified formula.

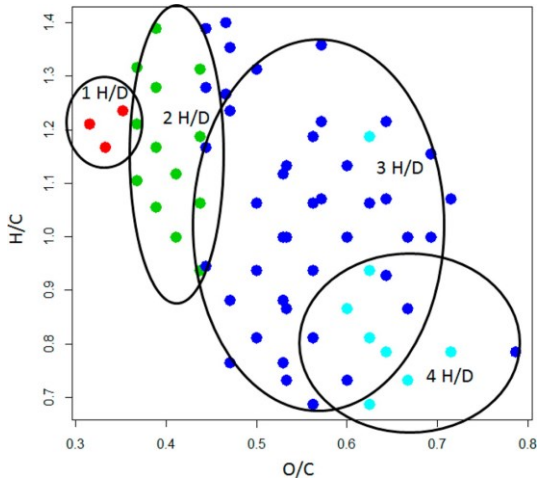


Figure 1.28 van Krevelen diagram for FTICR MS of SRFA labelled with D_2O spray. Circle/coloured regions represent number of H/D exchanges identified for each formula.[89] Reprinted with permission from Anal. Chem., 85/22, Kostyukevich. Y., Kononikhin. A., Popov. I., Kharybin. O., Perminova. I., Konstantinov. A., and Nikolaev. E., Enumeration of Labile Hydrogens in Natural Organic Matter by Use of Hydrogen/Deuterium Exchange Fourier Transform Ion Cyclotron Resonance Mass Spectrometry, 11007-11013. Copyright (2013) American Chemical Society.

Trimethylsilyldiazomethane (TMS-diazomethane) was discovered to be a safer, less volatile and easier to handle than diazomethane. TMS-diazomethane releases minimal diazomethane *in situ* and selectively methylates carboxylic acid groups, identically to diazomethane.[90] Studies of the mechanism of the reaction found that by including deuterated solvents it is possible to selectively label carboxylic acid groups with CD_3 groups.[91] These CD_3 methyl esters are suitable for analysis by FTICR MS due to the deuterons being found natively at detectable concentration in HS.

Chapter 2: Aims

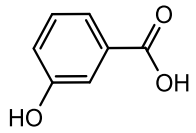
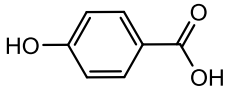
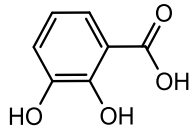
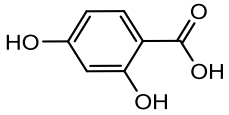
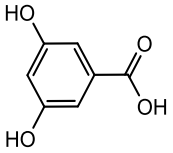
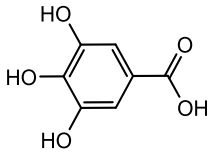
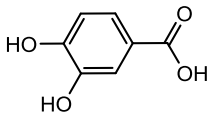
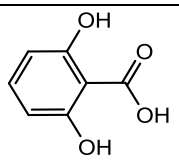
Due to the highly complex nature of humic substances, relating chemical properties to geochemical function is not yet possible today. This is mainly because the chemical structures of HS compounds are largely unknown. High-resolution NMR spectroscopy and FTICR MS methods are the most promising tool to structurally characterise these complex molecules, however, methodological advances are needed to unlock their potential. The aim of this work is to enrich the spectroscopic/spectrometric toolbox, in order to increase the volume and quality of information gathered. Specifically I aim to:

- Test and evaluate the usability and reliability of software for automated analysis of FTICR MS data
- Create new software tools for analysis of chemically modified humic substances
- Test the usability of Laser Desorption/Ionisation and contrasting it with electrospray ionisation, the standard ionisation method used in this field; and comparing the nature of chemical species ionised by both methods
- Investigate experimental limits of FTICR MS with regards to sample aggregation and fragmentation
- Analyse low and high mass species normally ignored by standard FTICR MS analyses of HS
- Perform selective methylation of carboxylic acid groups with isotopically enriched CD_3 functionality for enumeration at the level of individual molecules using FTICR MS
- Test nonselective methylation of labile proton functionalities on model mixtures with ^{13}C enriched methyl groups for study by NMR.
- Perform nonselective methylation of all labile proton functionalities with ^{13}C enriched methyl groups for NMR studies of HS

Chapter 3: Materials and Methods

3.1 Materials and model compounds

Table 3.1 Model compounds used in reaction testing with structure, mass and reference number.

Name	Structure	#	Monoisotopic Mass (Da)
3-hydroxybenzoic acid		**1**	138.03170
4-hydroxybenzoic acid		**2**	138.03170
2,3-dihydroxybenzoic acid		**3**	154.02661
2,4-dihyrdroxybenzoic acid		**4**	154.02661
3,5-dihydroxybenzoic acid		**5**	154.02661
3,4,5-trihydroxybenzoic acid		**6**	170.02153
3,4-dihydroxybenzoic acid		**7**	154.02661
2,6-dihydroxybenzoic acid		**8**	154.02661

SRFA was supplied by the International Humic Substances Society (IHSS).

3.2 TMS-diazomethane methylation

3.2.1 TMS-diazomethane methylation of model compounds

All glassware was washed with MeOH prior to use. Approximately 0.25 mmol (30 – 40 mg) starting material (models **1** – **7**) dissolved in 6.5 mL solvent solution, containing 20 % MeOH (1.3 mL) and 80 % toluene (5.2 mL). The solution is stirred for 1 hour before being transferred to an inert N_2 environment. 1 mmol (0.15 mL) TMS-diazomethane added followed by a further 30 minutes stirring. TMS-diazomethane adds a yellow/green colour to the solution, as the reaction progresses N_2 bubbles are evolved and the yellow/green colour of the solution gradually fades. After 30 minutes the reaction is quenched by dropwise addition AcOH until a colourless solution is achieved. The solvents are removed by vacuum evaporation, using a pressure of 40 mbar and at room temperature, temperature increased to 40 °C once the initial MeOH has been removed.

3.2.2 TMS-diazomethane deutero-methylation of model compounds

All glassware washed with MeOD prior to use. Approximately 0.25 mmol (30 – 40 mg) starting material (models **1** – **7**) dissolved in 5 mL MeOD, stirred for 15 minutes and dried under vacuum. This step is repeated a second time to allow for more complete exchange of labile protons for deuterons. The starting material is then dissolved in 6.5 mL solvent solution, containing ~13 % MeOD (0.8 mL) and ~87 % toluene (5.2 mL). The solution is stirred for 1 hour in an inert N_2 environment. 1 mmol

(0.15 mL) TMS-diazomethane is mixed with 0.5 mL MeOD and stirred for 30 minutes, the TMS-diazomethane solution is then added to the starting material solution followed by a further 30 minutes stirring. Experimental observations identical to section 3.2.1. After 30 minutes the reaction is quenched by dropwise addition AcOH until a colourless solution is achieved. The solvents are removed by vacuum evaporation, using a pressure of 40 mbar and at room temperature, temperature increased to 40 °C once the initial MeOD has been removed. The product is dissolved in 0.5 mL MeOH, stirred for 15 minutes and evaporated under vacuum, this is repeated two more times to encourage complete back exchange of the labile deuterons for protons.

3.2.3 TMS-diazomethane methylation of SRFA

The reaction was scaled down to prevent wasting the more precious HS samples. All glassware washed with MeOH prior to use. 10 mg SRFA is dissolved 1.3 mL MeOH, stirred for 15 minutes and dried under vacuum. This is repeated twice and is preformed to imitate the proton/deuteron pre-exchange of the deuteron-methylation and make these two sample truly comparable. The SRFA was then dissolved in 1.6 mL toluene. The solution is stirred for 1 hour before being transferred to an inert N_2 environment. 0.1 mL TMS-diazomethane is mixed with 0.4 mL MeOH and stirred for 30 minutes, the TMS-diazomethane solution is then added to the starting material followed by a further 30 minutes to 1 hour stirring. TMS-diazomethane adds a yellow/green colour to the solution, as the reaction progresses N_2 bubbles are evolved and the yellow/green colour of the solution gradually fades. After 30 minutes and once the reaction has lost most of the yellow/green colour the reaction is quenched by dropwise addition AcOH

until a colourless solution is achieved. The solvents are removed by vacuum evaporation, using a pressure of 40 mbar and at room temperature, temperature increased to 40 °C once the initial MeOH has been removed. The product is dissolved in 0.5 mL MeOH, stirred for 15 minutes and evaporated under vacuum, this is repeated two more times. The product is dissolved in H_2O, agitated for 1 hour and then dried by lyophilisation, this is also repeated two additional times to match the back exchange steps in the deuteron-methylation reaction.

3.2.4 TMS-diazomethane detuero-methylation of SRFA

The reaction was scaled down to prevent wasting the more precious HS samples. All glassware washed with MeOD prior to use. 10 mg SRFA is dissolved 1.3 mL MeOD, stirred for 15 minutes and dried under vacuum. This step is repeated a second time to allow for more complete exchange of labile protons for deuterons. The SRFA was then dissolved in 1.6 mL toluene. The solution is stirred for 1 hour before being transferred to an inert N_2 environment. 0.1 mL TMS-diazomethane is mixed with 0.4 mL MeOD and stirred for 30 minutes, the TMS-diazomethane solution is then added to the starting material followed by a further 30 minutes to 1 hour stirring. TMS-diazomethane adds a yellow/green colour to the solution, as the reaction progresses N_2 bubbles are evolved and the yellow/green colour of the solution gradually fades. After 30 minutes and once the reaction has lost most of the yellow/green colour the reaction is quenched by dropwise addition AcOH until a colourless solution is achieved. The solvents are removed by vacuum evaporation, using a pressure of 40 mbar and at room temperature, temperature increased to 40 °C once the initial MeOD has been removed. The product is dissolved in 0.5 mL MeOH, stirred for 15 minutes and evaporated under vacuum,

this is repeated two more times. The product is dissolved in H_2O, agitated for 1 hour and then dried by lyophilisation, this is repeated two more times to encourage complete back exchange of the labile deuterons for protons.

3.3 Methyl iodide methylation of HS

NaH (0.13 g) weighed out, attached to a Schlenk line and placed under an inert gas (N_2) atmosphere. 5 mL hexane added and the solution allows to stir gently for 5 minutes. Hexane was carefully removed using a long syringe needle and added to a beaker of waste MeOH to destroy any residual NaH, this leaves a dried NaH power (highly flammable with water). The hexane wash was repeated a second time. 10 mL anhydrous DMF added to the reaction vessel and stirring increased to moderate pace. Solution cooled to 5 °C using an ice bath and ~60 mg HS added. Reaction allowed to warm to room temperature with continual stirring for 1 h. 0.26 mL MeI (^{12}C or ^{13}C dependent on where the reaction is unlabelled/labelled respectively) added dropwise following which the reaction vessel is sealed. The solution is kept under inert atmosphere and stirring for 72 h. 15 mL $CHCl_3$ were added at the solution stirred for a further 5 h. Reaction was quenched using 20 mL H_2O and dropwise addition of concentrated HCl until the solution has a pH ~7. Reaction was worked up using $CHCl_3$/H_2O washes, followed by the organic layer being dried over $MgSO_4$ and filtered. Vacuum evaporation was performed at 60 °C and 37 mbar to remove the bulk of the solvent followed by 60 °C and 16 mbar to remove residual DMF. Dried products dissolved in 550 μL $CDCl_3$ for analysis by NMR.

3.4 NMR spectroscopy

3.4.1 NMR Instrumentation

NMR spectra were collected at the University of Edinburgh. Instruments used were a Bruker ADVANCE III HD 500 MHz (PRO500) spectrometer with a N_2 cooled Prodigy cryo-probe, tuned for ^{1}H and X nuclei; a Bruker ADVANCE III 500 MHz (AVA500) spectrometer with a He cooled DCH cyro-probe, tuned for ^{13}C (inner coil) and ^{1}H (outer coil); a Bruker ADVANCE III 800 MHz (AVA800) spectrometer with a He cooled TCI cryo-probe, tuned for ^{1}H (inner coil) and ^{13}C (outer coil).

3.4.2 NMR sample preparation

NMR samples were prepared by dissolving 10 – 20 mg of sample in 550 μL $CDCl_3$ or DMSO-d_6, followed by centrifugation for 5 minutes at 6.9 xG (8500 rpm) to remove undissolved particular matter. Sample solutions were transferred to an NMR tube for analysis.

3.4.3 NMR experiments

Experiments were performed using the PRO500 and AVA500 in automation mode using IconNMR and TopSpin (Bruker). Experiments performed using the AVA800 in manual mode using TopSpin.

3.5 FTICR mass spectrometry

3.5.1 FTICR Materials and Instrumentation

LC-MS grade methanol and LC-MS grade water were purchased from Fischer Scientific and used in the preparation of all MS samples.

FTICR mass spectra were collected at the University of Edinburgh on a 12 T SolariX FTICR MS (Bruker Daltonics), using *ftmscontrol* 2.1, coupled with ESI and MALDI, fitted with a solid-state 1 kHz smartbeamTMII laser, sources. All analyses were performed using negative ion mode. Data initially processed with *DataAnalysis* 4.2.

3.5.2 ESI-(-) method for software development and assignment tests

Standard method used to test daily performance of instrument. SRFA solution at 0.1 mg/mL freshly prepared in 50 % $MeOH_{(aq)}$. 200 scans were summed per FID. 200 µL/h flow rate of sample was used. Nebuliser pressure was set to 2 Bar, dry gas flow was set to 4 L/min at 180 °C. Low and high mass limits were set to *m/z* 110.6 and *m/z* 2000, respectively. Ions were accumulated for 700 ms with a collision cell voltage of 0.5 V. Transfer optics were set to 4 MHz frequency at an amplitude of 440 Vpp with a time of flight of 0.7 ms. ICR trapping potentials were -0.4 V (front) and -0.5 V (back). A 4 MW FID was collected.

3.5.3 ESI-(-), LDI-(-) and MALDI-(-) methods for comparison of ionisation techniques in the analysis of HS

Three samples were prepared for FTICR MS analysis. For MALDI-FTICR MS a 2 mg mL-1 SRFA/10 mg mL-1 DHB solution and for LDI-FTICR MS a 2 mg mL-1 solution in 50 % methanol:water solution were prepared. 1 μL of each solution was spotted onto a MTP 384 polished steel plate and dried at room temperature. For ESI-FTICR MS a 0.1 mg mL^{-1} SRFA solution in 50 % methanol:water was used.

For ESI data acquisition, a continuous flow sample was infused with a syringe flow rate of 200 μL h^{-1}. The nebuliser gas pressure was set to 2.0 bar. The drying gas was run at 180 °C and 4 L min^{-1}. The broadband spectra were acquired between *m/z* 150 and *m/z* 1000 in 2 MW FIDs and summed over 50 scans. Each scan had an ion accumulation time of 700 ms and a time of flight of 0.7 ms. For MALDI and LDI data acquisition the laser power was set to the minimum required to produce ions. Broadband spectra were acquired between *m/z* 150 and *m/z* 1000 in 2 MW FIDs and summed over 50 scans; selective accumulation was used excluding scans with low ion counts. Each scan had 1000 laser shots and an ion accumulation time of 500 ms. FIDs were zero filled once prior to Fourier transformation with the default processing parameters.

3.5.4 ESI-(-) methods used in the investigation of instrument parameters and mass bias on FTICR MS analysis of HS

3.5.4.1 General parameters

All experiments were acquired with the following settings unless otherwise noted. All FIDs were processed, including apodization and zero-filling, with default parameters of full sine apodization and zero-filling once prior to fast Fourier transform. 200 scans were summed into one FID. The ESI parameters were a flow rate of 200 µL/h, nebuliser pressure of 2 bar, dry gas flow of 6 L/min and temperature of 180°C. Source optics were as follows; capillary exit was set to -200 V, deflector plate was set to -180 V, funnel 1 was set to -100 V and skimmer 1 was set to -45 V. Funnel RF amplitude was set to 120 Vpp. The octopole was set to a frequency of 5 MHz with RF amplitude of 300 Vpp. The quadrupole Q1 mass was set to *m/z* 100. The ICR cell setting were as follows; a sidekick potential of 2 V, sidekick offset of 2 V, analyser entrance was set to 4 V, front trapping potential was -0.5 V and back trapping potential was -0.6 V. The 'normal' dataset was acquired with a 2 MW FID and low mass limit of *m/z* 147.44 and high mass limit of *m/z* 1000. Ions were accumulated in the collision cell for 150 ms with a voltage of 0.5 V and DC extract bias of -0.7 V, gas control was set to 35%, RF frequency of 2 MHz and collision RF amplitude of 400 Vpp. Transfer optics were set to a time of flight of 0.7 ms, frequency of 4 MHz and RF amplitude of 440 Vpp.

3.5.4.2 High concentration ESI-(-) analysis

Spectra were acquired at 1 mg/mL, 2.5 mg/mL and 5 mg/mL. At 5 mg/mL spray stability began to suffer. These data were acquired at 4 MW with a low mass limit of *m/z* 98.3 and high mass limit of *m/z* 2000. At higher concentrations, the ion

accumulation time (IAT) was severely reduced to 10 ms to prevent overfilling the ICR cell and to continue acquiring high quality spectra.

3.5.4.3 Effect of collision voltage on SRFA

Spectra were acquired at the 2.5 mg/mL sample concentration, following the same method as previously stated, but using collision cell voltages of 0.5 V (typical acquisition voltage), 2 V, 4 V, and 6 V.

3.5.4.4 Low and high-mass analysis of SRFA

For acquisition of a low-mass optimised dataset, the transient was increased to 4 MW. IAT was 300 ms. Low mass was set to *m/z* 73.5 and high mass *m/z* 1000.

For acquisition of a high-mass optimised dataset, the transient was increased to 4 MW. IAT was 200 ms. Low mass was set to *m/z* 245.7 and high mass *m/z* 5000. The time of flight was increased to 1.0 ms. Q1 mass isolation was set to *m/z* 1000 with resolution of *m/z* 600.

3.5.5 ESI-(-) method used in the analysis of isotopically labelled deutero-methylated products

3.5.5.1 ESI-(-) analysis of CD_3 methylated model mixture

Model mixture containing methylated **1**, **2**, **4**, **5**, **6** and **7** prepared into a 1 mg/mL sample in 50 % $MeOH_{(aq)}$. 10 scans were summed per 4 MW FID. 500 µL/h flow rate of sample was used. Nebuliser pressure was set to 2 Bar, dry gas flow was set to 4 L/min at 180 °C. Low and high mass limits were set to *m/z* 110.6 and *m/z* 5000, respectively. Ions were accumulated for 400 ms with a collision cell voltage of 0.5 V.

Transfer optics were set to 4 MHz frequency at an amplitude of 440 Vpp with a time of flight of 0.5 ms. ICR trapping potentials were -0.4 V (front) and -0.5 V (back).

3.5.5.2 ESI-(-) analysis of CH_3 methylated SRFA

CH_3 methylated SRFA solution at 0.5 mg/mL prepared in 100 % MeOH. 100 scans were summed per 8 MW FID. 500 µL/h flow rate of sample was used. Nebuliser pressure was set to 2 Bar, dry gas flow was set to 4 L/min at 180 °C. Low and high mass limits were set to *m/z* 110.6 and *m/z* 5000, respectively. Ions were accumulated for 700 ms with a collision cell voltage of 0.5 V. Transfer optics were set to 4 MHz frequency at an amplitude of 440 Vpp with a time of flight of 0.7 ms. ICR trapping potentials were -0.4 V (front) and -0.5 V (back). Quadrupole isolation perform to create two spectral windows, the first used a Q1 mass of *m/z* 258 and Q1 resolution of *m/z* 107, the second used a Q1 mass of *m/z* 584 and Q1 resolution of *m/z* 445.

3.5.5.3 ESI-(-) analysis of CD_3 methylated SRFA at 8 MW

CD_3 methylated SRFA solution at 0.5 mg/mL prepared in 100 % MeOH. 200 scans were summed per 8 MW FID. 200 µL/h flow rate of sample was used. Nebuliser pressure was set to 2 Bar, dry gas flow was set to 4 L/min at 180 °C. Low and high mass limits were set to *m/z* 110.6 and *m/z* 2000, respectively. Ions were accumulated for 700 ms with a collision cell voltage of 0.5 V. Transfer optics were set to 4 MHz frequency at an amplitude of 440 Vpp with a time of flight of 0.7 ms. ICR trapping potentials were -0.4 V (front) and -0.5 V (back). Quadrupole isolation perform to create five spectral windows, Q1 mass and Q1 resolution parameters for each of window is given in Table 3.2.

Table 3.2 Q1 mass and resolution values using the in collection of windowed spectra for deuteron-methylated SRFA at 8 MW.

Q1 Mass (*m/z*)	Q1 Resolution (*m/z*)
280.5	63
341.8	46
399.4	63
455.0	47
540.0	120

3.5.5.4 ESI-(-) analysis of CD_3 methylated SRFA at 4 MW

CD_3 methylated SRFA solution at 0.5 mg/mL prepared in 100 % MeOH. 200 scans were summed per 4 MW FID. 200 µL/h flow rate of sample was used. Nebuliser pressure was set to 2 Bar, dry gas flow was set to 4 L/min at 180 °C. Low and high mass limits were set to *m/z* 110.6 and *m/z* 2000, respectively. Ions were accumulated for 700 ms with a collision cell voltage of 0.5 V. Transfer optics were set to 4 MHz frequency at an amplitude of 440 Vpp with a time of flight of 0.7 ms. ICR trapping potentials were -0.4 V (front) and -0.5 V (back). Quadrupole isolation perform to create five spectral windows, Q1 mass and Q1 resolution parameters for each of window is given in Table 3.3.

Table 3.3 Q1 mass and resolution values using the in collection of windowed spectra for deuteron-methylated SRFA at 4 MW.

Q1 Mass (*m/z*)	Q1 Resolution (*m/z*)
300	90
381	22
413	26
453	50
540.0	116

3.6 Figure production

NMR figures were produced using MestReNova 11.0.4 (Mestrelab Research) and TopSpin 3.2 (Bruker). FTICR MS figures were produced using Data Analysis 4.2 (Bruker), Origin 2016 (OriginLab), FTMSVisualization scripts designed by Kew *et al.*[92], and modified versions of these scripts.

Results and Discussion

Chapter 4: FTICR MS Data Handling

4.1 Exact Mass

FTICR MS has high enough mass accuracy that it can separate out peaks that have the same nominal mass. *E.g.* $[C_2H_5]^-$ and $[CH_1O]^-$ both have the same nominal mass of 29 g/mol. However, the exact masses vary slightly from the nominal mass (^{1}H nominal mass = 1 g/mol, exact mass = 1.0079 g/mol and ^{16}O nominal mass = 16 g/mol, exact mass = 15.9949 g/mol). So the exact mass for $[C_2H_5]^-$ and $[CH_1O]^-$ is actually 29.0397 and 29.0033 g/mol, respectively. This difference can be clearly resolved via FTICR MS and thus these compounds can be differentiated from one another (Figure 4.1). All compounds with different formulae can be separated in this way, although this excludes isomeric and isobaric compounds that will still overlap.

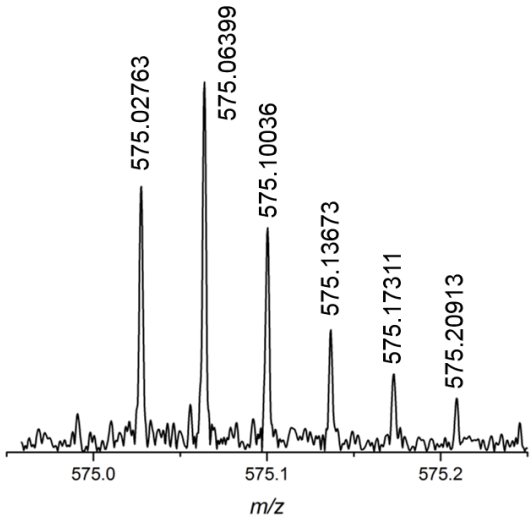


Figure 4.1 Peak distribution at *m/z* 575 for SRFA using negative ion ESI.

4.2 Formula Assignment and Kendrick Mass

With separation and mass accuracy it is then possible to assign formulae based on the exact masses. This can be done manually in most spectra manipulation programs by calculating all possible formulae for a nominal mass and then using the error between measured and calculated exact mass to determine which formula is correct. The formulae generated are all governed by an initial set of rules, either set by the program or defined by the user, these rules reflect possible elemental ratios such as a H/C value no greater than 4. The process of manual assignment in this manner takes a significant amount of time as a single sample can contain upwards of 3000 peaks. Therefore once a reasonable number of formulae have been assigned manually, Kendrick mass (KM) and Kendrick mass defect (KMD) can be used to speed up the assignment considerably. This can be done by graphically interpreting MS data. KM involves converting the mass of an atom or a small group to an integer value (*e.g.* C to 12.000 or CH_2 to 14.000) and then applying the same conversion to all masses (see Equation 9). Where exact mass is the exact IUPAC mass and nominal mass is the integer mass used to normalise against. KMD follows from KM and is the difference between the nominal mass and the KM of a formula (Equation 10).

$$KM = Exact\ Mass\ x\ \frac{Nominal\ Mass}{Exact\ Mass} \qquad (9)$$

$$KMD = Nominal\ Mass - Exact\ Kendrick\ Mass \qquad (10)$$

The outcome is a plot of KMD vs KM for every formula identified in the mass spectrum. The KMD values differ only very slightly, however they do form trends that can be easily seen when visualised graphically. For example, if the KM calculations

had been performed using CH_2 as the chosen molecular fragment, all formulae that differ by a multiple of CH_2 will have exactly the same KMD. Therefore, when KMD is plotted against nominal KM, horizontal lines occur formed by the homologous series of increasing CH_2 units. Displacement in the vertical axis allows for separation of different group types, *e.g.* cyan, blue and red in Figure 4.2 represent N, NO and NO_2, respectively. The repeating vertical displacement between the horizontal lines of the same group is caused by a difference in pairs of hydrogen atoms and number of double bonds.[93]

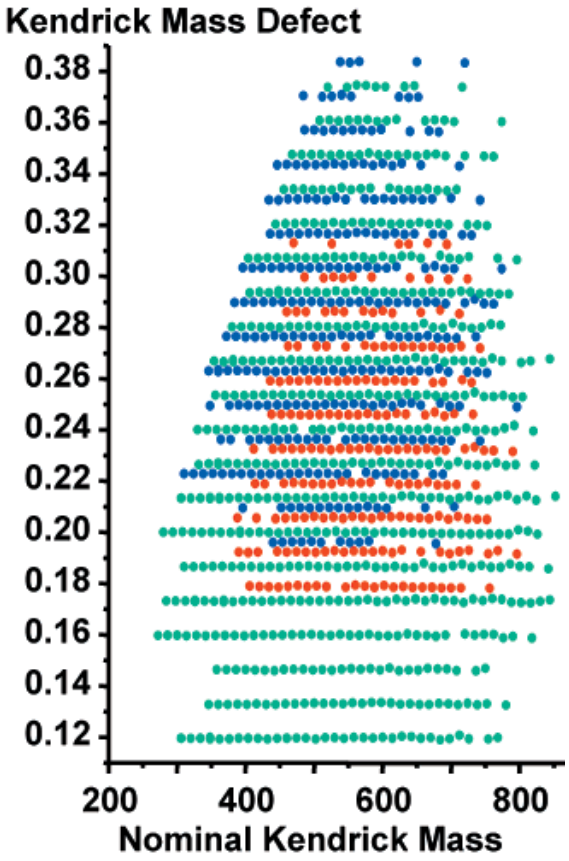


Figure 4.2 KMD vs KM showing CH_2 series horizontally and vertically N (cyan), NO (blue) and NO_2 (red).[93] Reprinted with permission from Anal. Chem., 73/19, Hughey. C.A., Hendrickson. C.L., Rodgers. R.P., and Marshall. A.G., Kendrick Mass Defect Spectrum: A Compact Visual Analysis for Ultrahigh-Resolution Broadband Mass Spectra, 4676-4681. Copyright (2001) American Chemical Society.

This can be used to aid the process for manual formula assignment of complex data sets as the chain of formulae for the same KMD all increase by the nominal KM species (usually CH_2). As a result, by identifying a small number of formulae for each KMD chain, the rest of the formulae present in that chain can be identified without need of the laborious manual assignment, as their formula is the same, plus or minus a certain number of KM units. For example, using the CH_2 KM series and starting on formula $C_xH_yO_z$ the next point along with a mass increase of 14 and the same KMD would belong to the formula $C_{(x+1)}H_{(y+2)}O_z$. This method can be used to assign most of the formulae in a chain quickly whilst only requiring a few thorough manual assignments. However this method could still potentially take hours to days, a timescale still viewed as too long. A multitude of software aimed at the reliable and rapid automated assignment of FTICR MS data was therefore proposed.

4.3 Automated Assignment

Due to the large data sizes generated from FTICR MS, especially for complex mixtures such as HS, manual assignment for every sample analysed is not feasible. Thus a method to quickly and reliably perform this assignment near automatically is needed before exploring other avenues of investigation. A large proportion of the software available is described as “in-house software”, for which only the authors themselves know the reliability and speed. Two different software choices were tested in this work, KMD_v.5 (David P. A. Kilgour) and PetroOrg s-10.2 (Florida State University).

4.3.1 KMD_v.5

KMD_v.5 is open source software that runs through LabView and assigns formulae using a two-step approach.[94] Initially a strict formula calculator comparison is perform, similar to manual assignment, where an error threshold for comparison is enforced. All formulae that fall within this error threshold are considered possible positive matches which can result in a single peak having several possible formulae. Only peaks that produce a single possible formula are considered reliably assigned. Following this, the software uses these reliably assigned formulae to start generating an interconnected map, based on the Kendrick mass analysis of the peaks. Trend lines for modifications to the formulae (+O, $-CO_2$, +C-OH, etc.) are extended beyond the limits of the map in order to detect peaks that were previously unassigned or multiply assigned. This allows the second step of the software to confirm identity of peaks that received multiple assignments and also search for unassigned peaks that were marginally outside error thresholds. I found that KMD_v.5 performs the task of automated assignment to an acceptable degree of accuracy (< 1 ppm) and reduces the time of assignment to a little under an hour, a vast improvement compared to the manual method. However it does suffer from a significant limitation that due to the intensity and scale of the calculations and algorithms employed the data sets are forced to be reduced in size in order to guarantee their correct and reliable performance. This upper limit is roughly 1500-2000 data points, which whilst not a small data set, is a significant reduction of the original size and risks misrepresenting and misanalysing the sample due to cutting approximately a third of the peaks. The software was tested against data from a SRFA standard sample, instrumental parameters given in Section 3.5.2, aiming to assign compounds containing C, H and O. From this 1009 unique formulae were assigned,

885 of which belong to monoisotopic formulae. The absolute error for assignment was 656 ppb. Whilst the average error is within a comfortable range for assigning CHO compounds (<1 ppm) combining monoisotopic and isotopic formulae this accounts for 33.2 % of the peaks present in the spectrum. The major reason for this low assignment levels is due to the aforementioned inference (second) step. Any peaks that cannot be assigned directly by the reasonably strict formula calculator are required to fit within the inference map, and peaks that produce many similar possible options or have too few connections within the map are dropped during this step. The former of these effects leads to lower assignment rates in highly complex samples like HS, where many formulae with only subtle accurate mass differences are seen across the entire mass range. The latter effect leads to a reduction in the number of ^{13}C isotope peaks being identified; this is due to the lower S/N ratio of ^{13}C peaks leading to the only inference connection being made to its monoisotopic partner. As a result this software was deemed unsuitable for use on HS samples due to its low assignment rate coupled with the necessity to exclude lower intensity peaks in order to achieve its stable performance. Kew *et al* have recently shown the negatives associated with trimming this type of data using their visualisation software as it results in false representation of the samples complexity and diversity, further confirming that KMD_v.5 is an unsatisfactory assignment tool for this study.

4.3.2 PetroOrg

The second piece of software I tested, PetroOrg, is a commercially available software developed for the assignment of formulae to complex petrochemical spectra. The software is not open source, therefore its exact algorithms are not known to the user.

It appears to work using a similar starting point to all other methods in that it uses a formula calculator to perform the major portion of assignment using user defined error thresholds and elemental limits. Then rather than inferring further formulae using KMD analysis, it checks all the assignments against one another, relating similar formulae by predefined groups such as similar heteroatom count, which it groups, this allows for the smaller and less reliable groups to be scrutinised upon. Due to the strictness of error thresholds a basic calibration, using a standard list, or using the on-line software calibration would give noticeably worse results. Thus it was required to perform a much more strict calibration using trusted formulae across the whole mass range and with a minimum spacing between points of 14 Da. However the payoff in effort for performing a significantly more thorough calibration is rewarded as the software will assign the full data set, to a high accuracy in roughly 10 seconds. Using the same SRFA spectrum as was used in the KMD_v.5 analysis a total of 2133 unique formulae were assigned, 1576 being monoisotopic and with an average absolute error of 304 ppb. This represents an assignment rate of 70.2 %, which is a vast improvement in comparison to the KMD_v.5 software with over double the assignment rate.

Whilst an assignment rate of ~70 % could be improved upon, due to the speed of the process, the reliability of assignment and the few thousand formulae that are able to be assigned, I concluded that PetroOrg is a suitable option for the task of HS assignment. Even considering the caveat of needing a more thorough calibration, it still outperforms the competing software and provides a suitable platform for reliable analysis. As such, all formula assignments were, from this point onwards, performed using PetroOrg.

4.3.3 Standard analysis of SRFA

In order to test the FTICR mass spectrometer for daily performance I decided to have a standard operational method and sample that could be run upon first using the instrument. Due to it being an international standard coupled with the thorough analysis already performed during automated assignment testing SRFA was selected as an appropriate standard for this task. The instrumental parameters for running this standard are unchanged from Section 4.3.1 and are described in Section 3.5.2. Figure 4.3 shows the broadband spectrum of SRFA run under these conditions. The data from previous standard acquisitions were compared to the standard data from that operation before data collection continued to make sure no significant change could be seen in the: signal intensity, number of peaks identified at each *m/z*, presence and influence of contaminants; presence of peak splitting; spray stability; and signal stability.

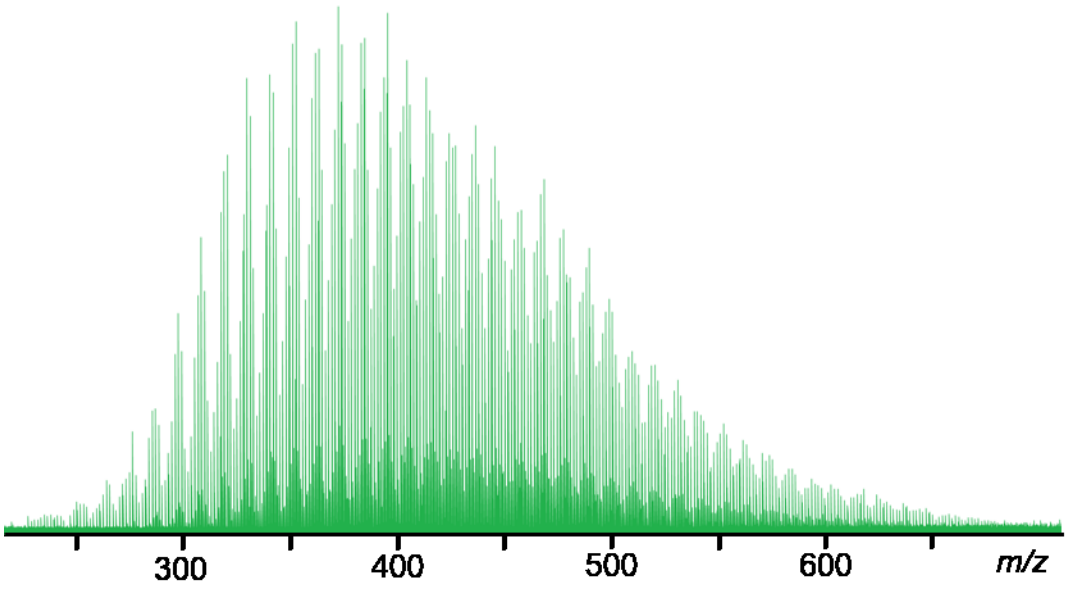


Figure 4.3 Negative ion ESI FTICR MS broadband spectrum of SRFA used to confirm instrument reliability.

This standard was used to produce an initial calibration list which could be used across all future standards to check instrument performance. The calibration list was built up, starting with the SRFA sample being spiked with ESI tuning mix, which gives two signals in the sample range at *m/z* 301.99 and *m/z* 601.99. Using one of these signals at a time a single point calibration was performed on the peak, allowing for reliable formula calculation within a 1 *m/z* range. For the calibration peak at 301.99 *m/z* the range from *m/z* 302.0 – 302.2 was used and the peaks were assigned using a formula calculator, the most reliable peak was selected as the one with the lowest assignment error, a matching ^{13}C isotope peak, and high resolution. A series of peaks following the CH_2 series from this original formula were identified until the peak intensity or resolution dropped too low. A shortened calibration list was then made from this series and used to calibrate a subsection of the spectrum with a calibration point every 14 *m/z* units. Within this calibrated region another CH_2 series was identified using a formula calculator and the series was followed beyond the current calibration range identical to the first series. This process was repeated until enough calibration points had been acquired to cover the full range of the spectrum, with a maximum separation of 14 *m/z* units. This process was repeated using the *m/z* 601.99 tuning mix signal in order to cross reference and confirm formula assignments, resulting in a basic calibration list (Table 4.1) that could be used on the SRFA and adapted as required for different samples.

Table 4.1 Basic calibration list for negative mode SRFA with ionic formulae that are deprotonated and carry a 1- charge with their exact mass in *m/z*

Ionic Formula	Exact Mass (*m/z*)	Ionic Formula	Exact Mass (*m/z*)
$C_{10}H_{11}O_8$	259.045941	$C_{25}H_{33}O_{12}$	525.197750
$C_{11}H_{13}O_8$	273.061591	$C_{26}H_{35}O_{12}$	539.213400
$C_{12}H_{15}O_8$	287.077241	$C_{23}H_{13}O_{16}$	545.020908
$C_{13}H_{17}O_8$	301.092891	$C_{24}H_{15}O_{16}$	559.036558
$C_{14}H_{19}O_8$	315.108541	$C_{25}H_{17}O_{16}$	573.052208
$C_{15}H_{21}O_8$	329.124191	$C_{26}H_{19}O_{16}$	587.067858
$C_{16}H_{23}O_8$	343.139841	$C_{27}H_{21}O_{16}$	601.083508
$C_{17}H_{25}O_8$	357.155491	$C_{28}H_{23}O_{16}$	615.099158
$C_{16}H_{19}O_{10}$	371.098370	$C_{29}H_{25}O_{16}$	629.114808
$C_{17}H_{21}O_{10}$	385.114020	$C_{30}H_{27}O_{16}$	643.130458
$C_{18}H_{23}O_{10}$	399.129671	$C_{31}H_{29}O_{16}$	657.146108
$C_{19}H_{25}O_{10}$	413.145321	$C_{32}H_{31}O_{16}$	671.161758
$C_{20}H_{27}O_{10}$	427.160971	$C_{33}H_{33}O_{16}$	685.177409
$C_{19}H_{21}O_{12}$	441.103850	$C_{31}H_{23}O_{19}$	699.083902
$C_{20}H_{23}O_{12}$	455.119500	$C_{32}H_{25}O_{19}$	713.099552
$C_{21}H_{25}O_{12}$	469.135150	$C_{33}H_{27}O_{19}$	727.115202
$C_{22}H_{27}O_{12}$	483.150800	$C_{34}H_{29}O_{19}$	741.130852
$C_{23}H_{29}O_{12}$	497.166450	$C_{35}H_{31}O_{19}$	755.146502
$C_{24}H_{31}O_{12}$	511.182100		

Chapter 5: Method development – Laser Desorption/ Ionisation for studies of humic substances

Matrix-free LDI is significantly less used in complex mixture analysis. As discussed in Section 1.3.3 this is due to an initial interpretation of fragmentation products being formed by LDI.[66] That did not fit with the understanding of HS being large polymeric molecules that was believed at the time.[69] However, subsequent studies have shown that HS are aggregates of small molecules.[15] This breakthrough in understanding the nature of HS raised the need for assessing viability of LDI as a possible ionisation method for this field.

5.1 SRFA, comparison of ionisation techniques

5.1.1 MALDI, LDI, and ESI FTICR mass spectra

SRFA, an IHSS standard, was selected as a suitable sample, and the performance of three ionisation techniques, MALDI, LDI, and ESI, was compared. Their superficial inspection highlights a high degree of similarity between the MALDI and LDI spectra (Figure 5.1), while the ESI spectrum is clearly different.

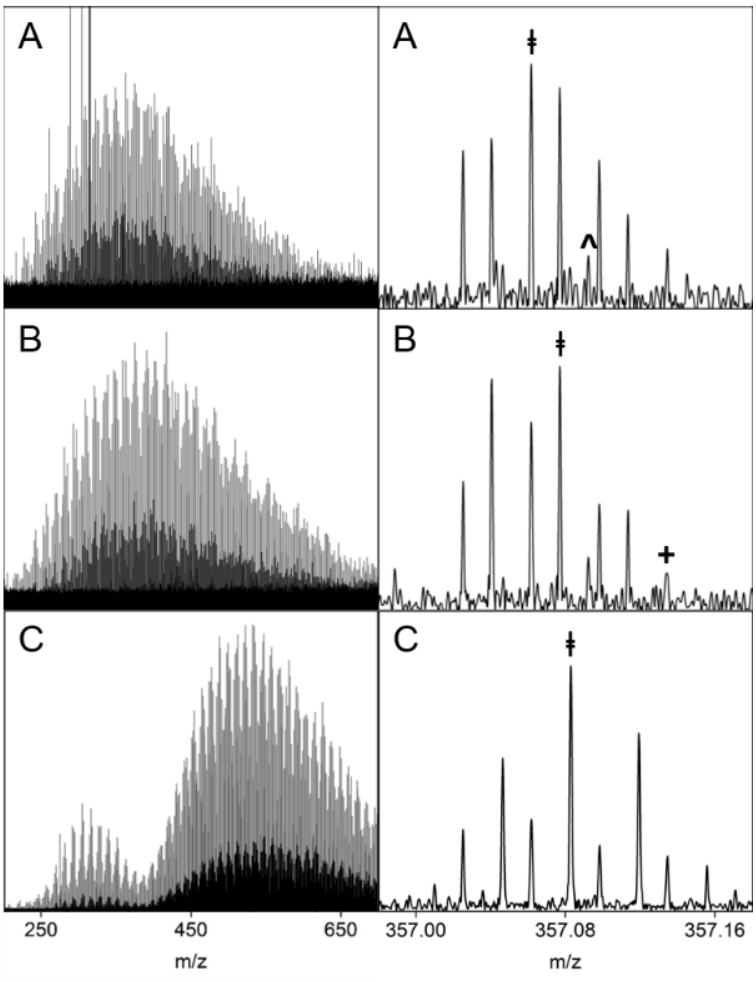


Figure 5.1 FTICR mass spectra of SRFA obtained by using MALDI (A), LDI (B), and ESI (C) ionisation methods. The left-hand spectra span the range of *m/z* 200−700; the right-hand spectra show peaks at the nominal mass of *m/z* 357. The peaks labelled with a double dagger were used for the S/N comparison. The highlighted peaks are present in both MALDI and LDI spectra but below the S/N threshold for MALDI (∧) or LDI (+).

For a quantitative comparison, 1396, 2209, and 5610 peaks were identified in each of the MALDI, LDI, and ESI spectra, respectively, over the displayed range of *m/z* values

using S/N threshold of 4. Out of these, it was possible to assign monoisotopic molecular formulae to 1046, 1450, and 2720 peaks, with average assignment errors of 102, 107, and 338 ppb, respectively. The average sigma values of assignments for peaks shown in the right panels of Figure 5.1 were 426, 317, and 368 mσ. Altogether, 43, 221, and 1464 peaks were assigned to molecular formulae containing one carbon-13 isotope. Cumulatively, this represents assignment of 78.0%, 75.6%, and 74.6% of the peaks present in each individual spectrum. Unassigned peaks either belong to real ions outside assignment error thresholds, examples of which can be seen in Figure 5.1, secondary carbon isotope peaks, noise, or signal processing artefacts. Figure 5.2 shows a distribution of 3815 unique molecular formulae identified from the MALDI, LDI, and ESI spectra of the SRFA. This figure also shows the count of molecular formulae common to diff erent combinations of ionisation methods.

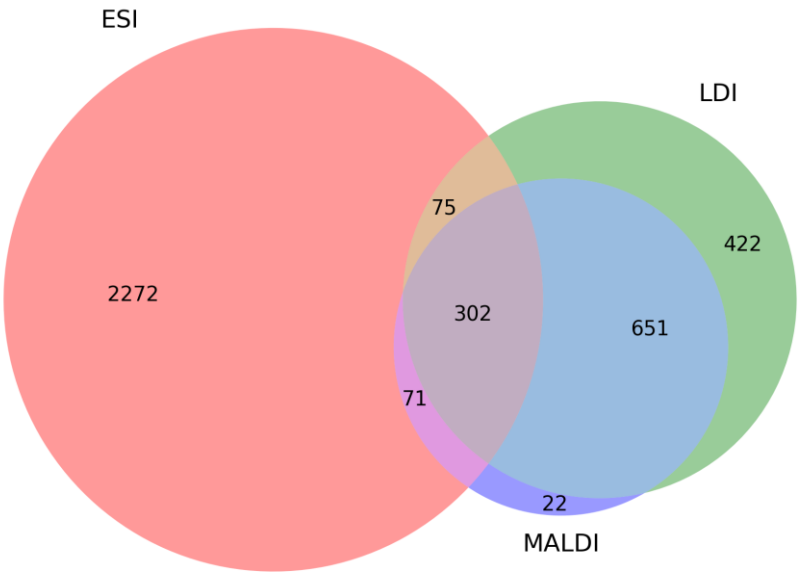


Figure 5.2 Venn diagram displaying the number of formulae identified that are unique to a single ionisation source or present from multiple sources using MALDI, LDI, and ESI.

It can be seen that MALDI and LDI spectra produced very similar data with over 90% matching formulae relative to the total number of peaks in the MALDI spectrum. Focusing on these two spectra, manual inspection of the nonmatching peaks revealed that their existence can be linked to one of three causes: (i) there is an absence of a peak in one spectrum; (ii) the same peak is present in both spectra, although one of the peaks falls outside of the 1 ppm assignment threshold, and hence, a formula is not assigned; (iii) peaks are present in both spectra, but in one data set, the signal is outside the peak picking parameters. These points are illustrated in Figure 5.1, where a region showing the peak distribution at 357 Da for all three ionisation methods is presented. Seven formulae were identified by all ionisation methods, and while visually the MALDI and LDI spectra contain the same peaks, one formula from each (labelled ∧ for MALDI and + for LDI) is below the S/N threshold. This led to a discrepancy between assignments as discussed above and was typical for odd *m/z* values in the 250−650 *m/z* range. Comparing all three methods, the S/N ratio was best for the ESI spectrum and poorest for the MALDI spectrum. For example, with reference to the double dagger labelled peaks in Figure 1, the S/N values were 21.8 (MALDI), 29.3 (LDI), and 39.3 (ESI). Although poorer than ESI, LDI is still substantially better than MALDI in this regard. An unavoidable problem in small molecule MALDI spectra is that the matrix peak is significantly larger than the sample peaks. This suppresses the sample peaks and reduces the average S/N in the whole spectrum; the improved quality of the LDI spectrum is the main reason why a greater number of molecular formulae could be assigned in the LDI spectrum (1450) compared to the MALDI spectrum (1046). The analysis presented thus far (Figures 5.1 and 5.2) suggests that there is a significant difference in the number of matching formulae between the molecules

ionised by ESI in comparison to MALDI/LDI. Hereafter, I explore these differences at the molecular level.

5.1.2 van Krevelen Diagrams of MALDI, LDI, and ESI Data

ESI is understood to ionise compounds that can exist as ions in solution;[95] therefore, it is primarily suited for observing protic compounds. MALDI/LDI, however, ionises compounds through different mechanisms[96] and, therefore, produces spectra that include less molecules with protic functionalities.

As HS are thought to be dominated by CHO compounds, it is reasonable to assume a correlation between the number of oxygens, polarity, and compounds ionised by ESI. This assumption is supported by an inspection of van Krevelen diagrams produced for MALDI, LDI and ESI spectra in Figure 5.3.

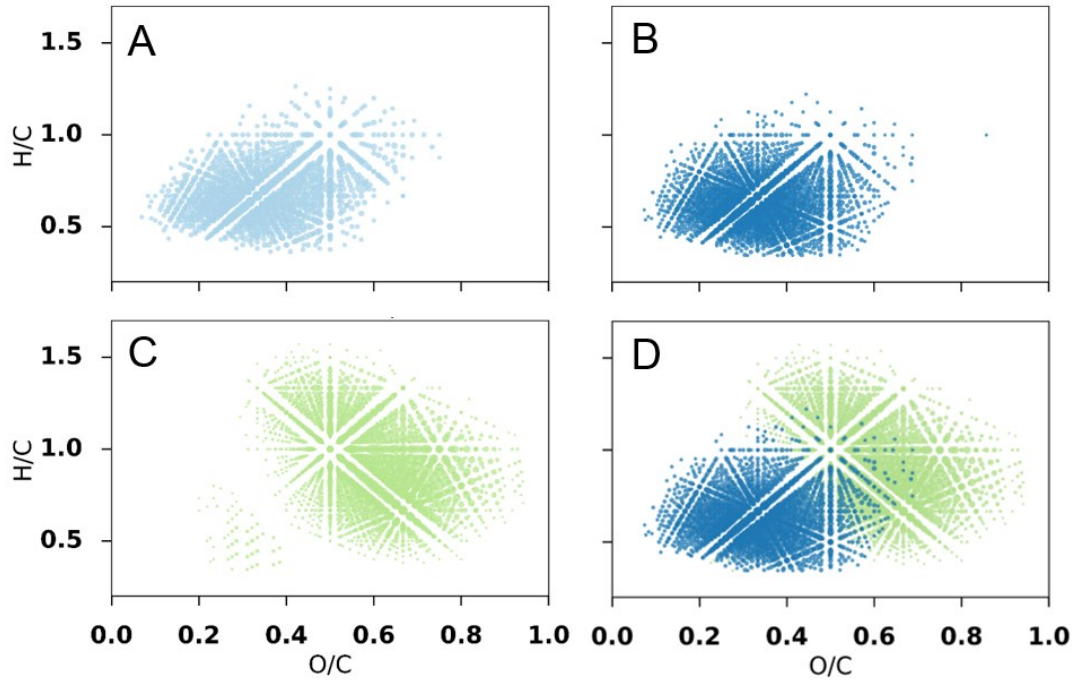


Figure 5.3 van Krevelen diagrams constructed using all formulae identified in the FTICR MS of SRFA: A, B and C display the datasets of MALDI, LDI and ESI, respectively, while D shows an overlay of the LDI (blue) and ESI (green) data sets emphasizing the significant difference in the type of compounds ionised by the two methods.

These diagrams clearly illustrate that MALDI/LDI and ESI ionise compounds with different O/C and H/C ratios. While LDI/MALDI ionise low O/C and low H/C compounds, the opposite is true for ESI. Taken together, the majority of compounds separate into the lower left (MALDI/LDI) and the upper right (ESI) regions of the van Krevelen plots, with an overlapping section in the centre.

5.1.3 Heteroatomic Class Distribution

The identified molecular formulae were further classified using their heteroatomic classes. A bar chart representing the count of O2 to O24 formulae in MALDI, LDI, and ESI FTICR MS spectra is shown in Figure 5.4. The results reinforce a similarity between MALDI and LDI spectra, which share high abundance of low oxygen compound classes. Their bell-shaped distribution has a maximum at the O8 class. A minimal number of O2−O4 compounds were assigned in the ESI spectra, with a maximum compound count shifted toward high oxygen classes and spread over a broader range of values. Here, MALDI/LDI were less effective and failed to ionise any compounds above the O16 class.

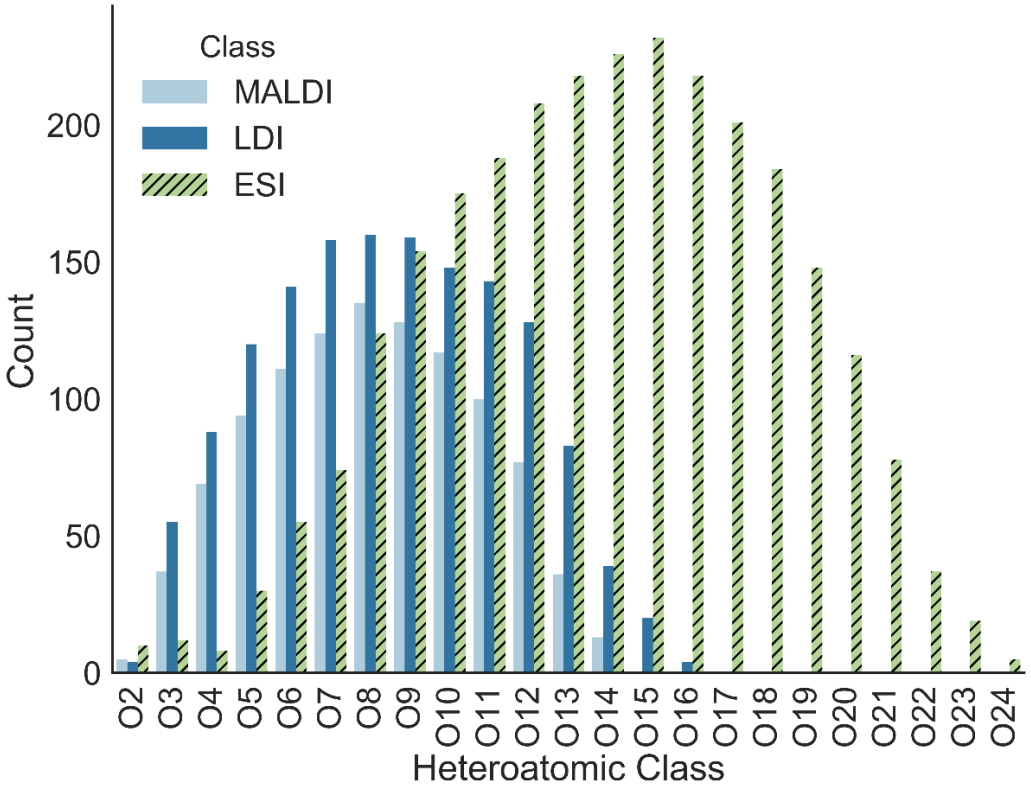


Figure 5.4 Bar chart of the number of formulae identified for oxygen classes from O2 to O24 for MALDI, LDI, and ESI FTICR MS spectra of SRFA.

5.1.4 Comparison of Molecular Formulae Assigned to LDI and ESI Spectra

Compounds with identical molecular formulae identified in both LDI and ESI spectra (377 or 9.9% of unique formulae to ESI and LDI; see Figure 5.2) cluster in the middle of the van Krevelen plot in Figure 5.3. It was initially assumed that these matching peaks mostly belong to the oxygen classes where the ionisation potential appears to be similar for both LDI and ESI, such as O9, where 49 of 154 (31.8%) formulae matched (see Figure 4). However, a specific comparison of oxygen classes showed a similar percentage of matching formulae for lower oxygen classes, *e.g.*, 14 of 55 (25.5%) for O6. Higher numbers of matches were found for higher oxygen classes, *e.g.*, 38 of 83 (45.8%) formulae matched in the O13 class. This observation suggests that the compounds ionised by LDI are significantly different to ESI compounds across the full range of oxygen classes and the matching formulae seen in the central part of the Figure 3 either belong to (i) identical molecules ionised by both LDI and ESI or (ii) structural isomers, each ionised only by one of the two methods. Due to the different ionisation mechanisms of LDI and ESI, the latter explanation is more likely.

5.1.5 Comparison of Molecular Formulae Assigned to MALDI and LDI Spectra

A similar comparison of the MALDI and LDI spectra categorized by the oxygen classes showed that the most populated classes contain most matching formulae, relative to the number of formula identified by MALDI (O7, 121 of 124 (97.6%), and O8, 124 of 135 (91.9%)) while the least populated classes showed fewer matching formulae (O3, 33 of 37 (89.2%), and O13, 28 of 36 (77.8%)). This fits with the

aforementioned observation that the major cause of discrepancy between the MALDI and LDI data sets is due to peaks dropping below the S/N threshold, an occurrence that is found at the extremes of the spectrum (lower *m/z* for low O classes and higher *m/z* for high O classes) where the S/N ratio is at its worst. It is therefore possible to conclude that identical compounds are ionised by both methods.

5.1.6 Aromaticity Index

The assigned molecular formulae were further analysed to characterize the aromaticity of molecules they represent. AI, unlike DBE, takes into account the presence of oxygen atoms in molecular formulae.[62] The amount of oxygen considered can however be reduced, assuming that not all oxygen atoms belong to carbonyl groups. As hydroxyl or ether oxygen is present in SRFA molecules alongside the carbonyl oxygen, a modified variant, AI_{mod}, which only counts half of the oxygen atoms and assumes the other half is sigma bound, was used. AI_{mod} can be defined as $AI_{mod} = (1 + C - 0.5O - S - 0.5H)/(C - 0.5O - S - N - P)$. The following threshold limits were set for AI_{mod} to categorize the aromatic character of the molecules: $AI_{mod} \leq 0.5$ for nonaromatic molecules, $0.67 \geq AI_{mod} > 0.5$ for aromatic molecules, and $AI_{mod} > 0.67$ for condensed aromatic molecules. Figure 5.5 shows stacked bars representing normalised percentages of formula counts that fall above and below the specified AI_{mod} thresholds.

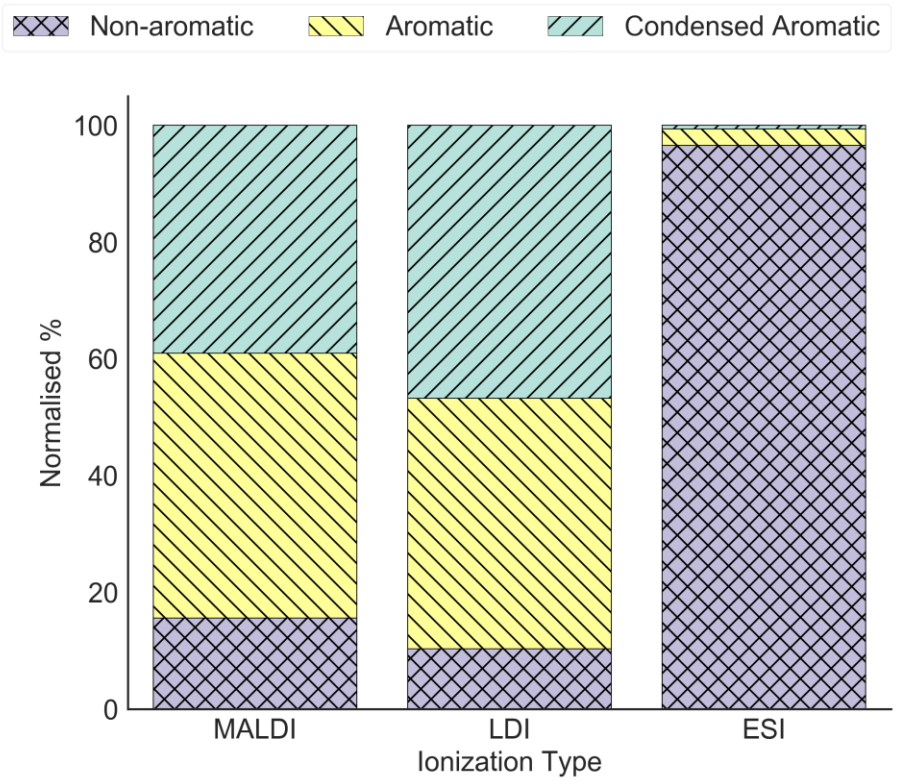


Figure 5.5 Stacked bar plot representing normalised % of formula count that belongs to three AI_{mod} categories; (i) $AI_{mod} \leq 0.5$ (nonaromatic, purple), (ii) $0.5 < AI_{mod} < 0.67$ (aromatic, yellow), and (iii) $AI_{mod} > 0.67$ (condensed aromatic, green) for MALDI, LDI, and ESI spectra.

This classification allows further interpretation of the character of molecules ionised by each method. It appears that LDI ionised the largest amount of compounds classified as condensed aromatics (46.8%), more than MALDI (39.0%). This is at the expense of nonaromatic compounds, while the relative amount of aromatic compounds ionised remained approximately constant (42.9% vs 45.4%) between the two techniques. To the contrary, 96.6% of compounds ionised by ESI could be classed as nonaromatic.

5.2 SRNOM, comparison of ionisation tehcniques

5.2.1 LDI and ESI FTICR MS of SRNOM

SRNOM, another IHSS standard, was used to additionally examine the performance of LDI and ESI as well as a test for a different type of complex organic matter. For LDI 1462 monoisotopic formulae and 98 single carbon isotope formulae were identified from a total of 2285 peaks. For ESI 1769 monoisotopic formulae and 383 single carbon isotope formulae were identified from a total of 2268 peaks. This respectively represents assignment rates of 68.3 % and 94.9 %. Of the monoisotopic formula assigned, 121 were common to both LDI and ESI (Figure 5.6), 3.9 % of the collective unique formula.

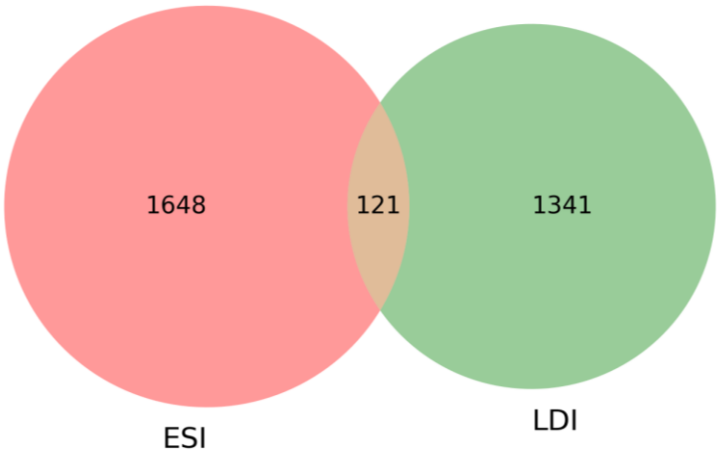


Figure 5.6 Venn diagram displaying the number of formulae identified unique to ESI, LDI and common between the two.

5.2.2 Heteroatomic class distribution

Similarly to Section 5.1.3 the formulae were classified based on heteroatomic class with reference to oxygen, ranging from O1 to O23. For SRNOM the difference in formula distribution is even more apparent, with LDI ionising several hundred species in the O1 – O5 distribution, while ESI failed to ionise compounds from this region, Figure 5.7. The maximum ionisation potential for LDI and ESI can be seen at O8 and O15, respectively. These observations are similar to those for SRFA, however, show even larger ionisation discrepancy, with ESI preferentially ionising higher oxygen number classes. As observed for SRFA, LDI performs in the exact opposite manner by preferentially ionising low oxygen number classes and tailing off for high oxygen number classes.

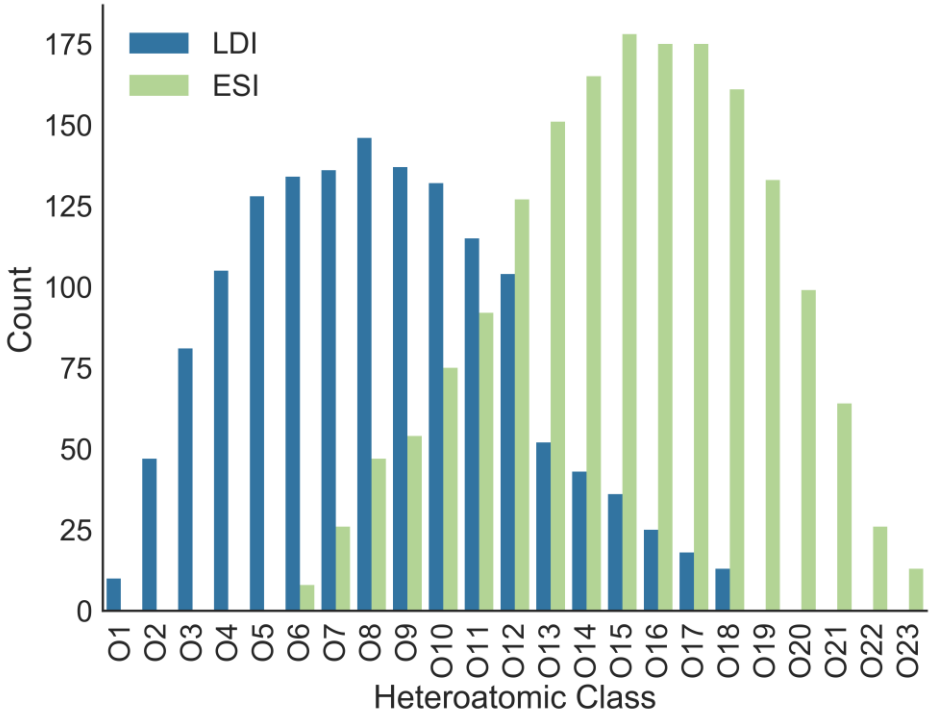


Figure 5.7 Bar chart of the number of formulae identified for oxygen classes from O1 to O23 for LDI and ESI FTICR MS spectra of SRNOM.

5.2.3 Aromaticity Index

Similarly to SRFA, the aromaticity of the sample and the species ionised were clarified using AI_{mod}. The same calculations and thresholds as used in Section 5.1.6 were applied to the SRNOM data sets and the normalised percentages of non-aromatic, aromatic and condensed aromatic can be seen in Figure 5.8.

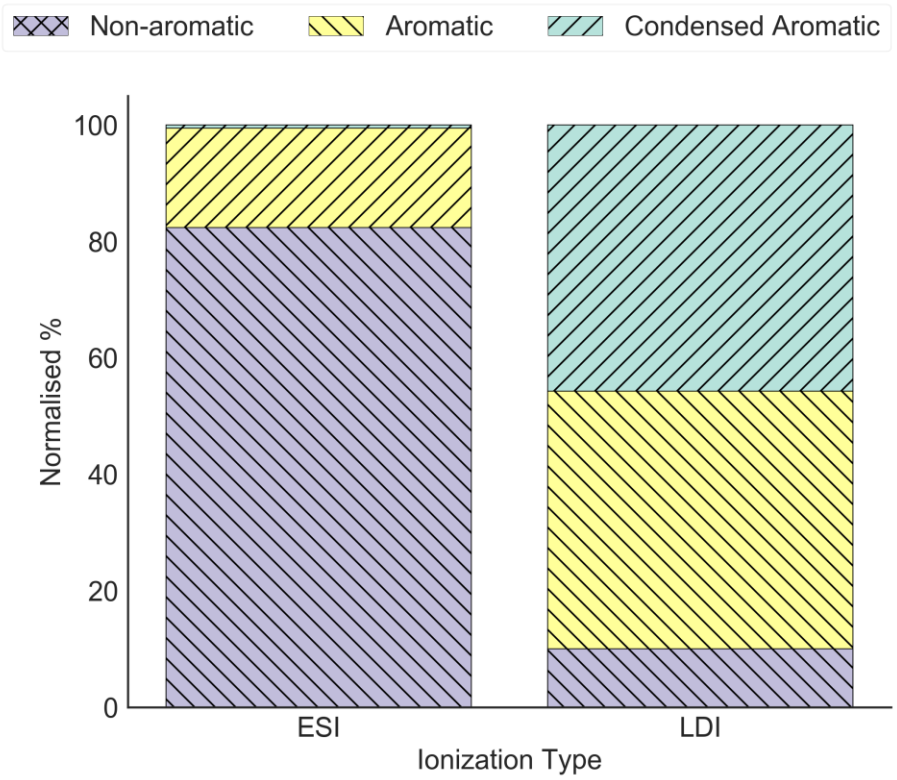


Figure 5.8 Stacked bar plot representing normalised % of formula count that belongs to three AI_{mod} categories; (i) $AI_{mod} \leq 0.5$ (nonaromatic, purple), (ii) $0.5 < AI_{mod} < 0.67$ (aromatic, yellow), and (iii) $AI_{mod} > 0.67$ (condensed aromatic, green) for LDI and ESI spectra of SRNOM.

Similarly to SRFA, ESI of SRNOM ionised mostly non-aromatic compounds (67.7 %), however it ionised proportionally more aromatic and condensed aromatic compounds (32.3 %) in comparison to SRFA (3.4 %). This suggests that SRNOM may contain more protic aromatic moieties in comparison to SRFA. LDI again ionised significantly more aromatics and condensed aromatics than ESI, 41.5 % and 42.8 %, respectively. This was more similar to the analogous statistics for SRFA (46.8 % and 42.9%).

5.3 Implications for MS studies of NOM

Hertkorn *et al*.[64] compared ESI, APCI, and APPI spectra of SRFA in both positive and negative modes. Due to the limited assignment capabilities available at the time of the study, molecular formulae were assigned only to 15−25% of the observed peaks. Among these, only 1.7% and 3.8%, for the positive and negative mode, respectively, were identical for all three ionisation techniques. In this study, 68−94% of observed peaks were assigned a molecular formula. This increase in reliable assignment was achieved through the work discussed in Section 4. Analysis of the data based on van Krevelen diagrams, heteroatomic class distributions, and aromaticity indices for both SRFA and SRNOM showed that distinct types of compounds are ionised by ESI and LDI. All three studies therefore demonstrated that ESI alone does not off er a complete coverage of NOM molecules. Importantly, a close match (>90%) was seen in this study between the molecular formulae identified in MALDI and LDI spectra of SRFA. While MALDI is an established method, it sees much less use in the investigation of NOM than ESI. As illustrated here, LDI, an even less used method, produced spectra superior to the MALDI method, ionising identical compounds while avoiding some of the

caveats of MALDI. The sample preparation for LDI is significantly simpler than for MALDI, as it does not require a trial and error process of selecting a matrix and its concentration. The complexity of HS makes the sample itself act as a matrix, enabling ionisation. It is therefore a simple "spot and shoot" method. Overall, the results challenge the dogmatic assessment of LDI as an inappropriate ionisation method for MS investigation of NOM.

FTICR MS spectra of NOM are today mostly acquired using an ESI as the preferred ionisation method. Significant differences between the ESI and LDI FTICR MS spectra of SRFA and SRNOM observed in this study complement differences seen between other ionisation techniques[64] and endorse the view that no single method is able to ionise all NOM compounds. To maximize the coverage by FTICR MS of the molecular space occupied by these complex mixtures, multiple ionisation methods must therefore be used. As a particularly convenient and readily deployable ionisation technique, LDI should therefore be included in standard analytical protocols for FTICR MS analysis of NOM.[97]

5.4 Further investigation into ESI-(-) MS spectra of SRFA

5.4.1 Comparison of 'normal' and bimodal spectra

Upon further inspection of the SRFA ESI-(-) spectrum, Figure 5.1, an interesting side note was observed that the data had an apparent bimodal distribution (*m/z* 200-400 and *m/z* 400-800), and that neither part of the distribution was centred at around *m/z* 350 where the SRFA typically displays a normal distribution, extending between *m/z* 200 and *m/z* 700.[79, 98, 99] In order to investigate this further I acquired an additional ESI-(-)

spectrum of SRFA using the same conditions as before, but tuned the spectrometer to produce a "normal" distribution. This newly acquired FTICR MS spectrum of SRFA (Figure 5.9a; see Section 3.5.4 for experimental parameters) extends from 200 to 1000 *m/z*, with a maximum around 380 *m/z*, and is thus more similar to the 'normal' spectrum as seen in Figure 3.3 which spanned the 200-700 *m/z* range, with a maximum at 300-350 *m/z*. A direct comparison of the entirety of the bimodal data with the new "normally distributed" data would be affected by differences in the signal-to-noise ratios (SNR), influencing peak picking and therefore assignments made. Thus, the subset region of 250-600 *m/z* was chosen to compare assignments, as this region was not at the limits of *m/z* values for either dataset. Within this region, the new spectrum allowed 1493 monoisotopic assignments (the bimodal spectrum had 1474), and 1282 of these were common to both. This 86% commonality of assignments shows that the data acquired is, broadly, the same. The differences observed are primarily due to the different SNR, especially at around 420 *m/z*, where one spectrum has a maximum intensity and the other a local minimum. To put these results into perspective, Kido Soule *et al.* found that technical replicates of ESI-(-) FTICR MS spectra of SRFA achieved a maximum of 87% peak commonality in their methodology.[100] Additionally, Sleighter *et al*. investigated intra-day variations by acquiring duplicate and triplicate of complex samples, achieving peak commonality of 75% on average for a SNR threshold of 5.[101] It is therefore reassuring to observe this level of agreement between the spectra across months and for apparently dissimilar distributions profiles.

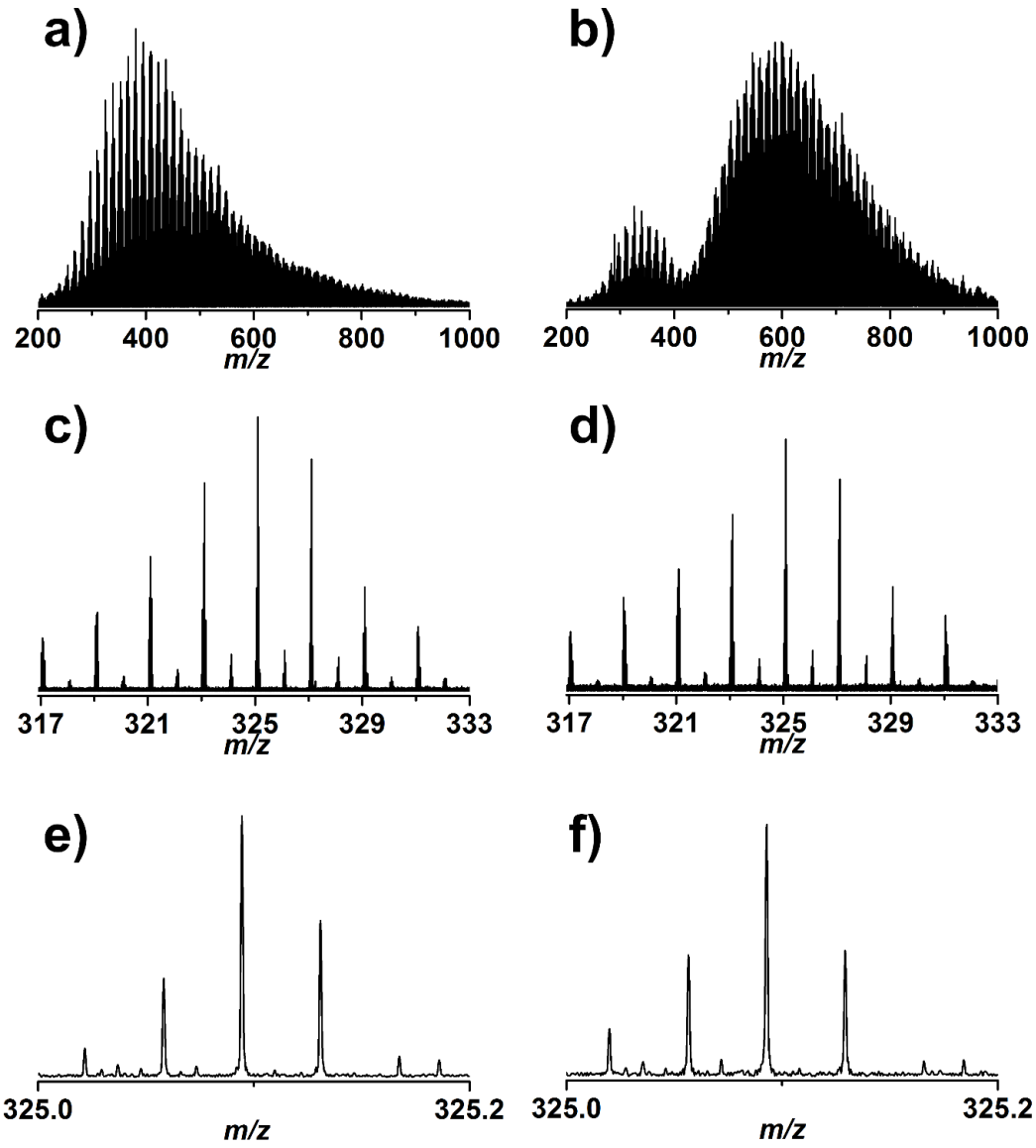


Figure 5.9 Negative mode ESI FTICR mass spectra of SRFA. Left column shows the broadband normally distributed data (a), a per-14 *m/z* expansion for this data (c), and a single nominal mass, 325 *m/z*, for this data (e). Right column shows the broadband bimodal data (b), the per-14 *m/z* expansion for the bimodal dataset (d), and a single nominal mass for the bimodal data (f).[102] Reprinted with permission from Anal. Chem., 90/9, Kew. W., Blackburn. J.W.T., and Uhrin. D., Response to Comment on "Laser Desorption/Ionization Coupled to FTICR Mass Spectrometry for Studies of Natural Organic Matter, 5968-5971. Copyright (2018) American Chemical Society.

To further emphasise the similarities between these two data sets, heteroatomic class distributions are shown for the same subset mass range (*m/z* 250-600) for both spectra in Figure 5.10a. Here, the distribution is effectively identical. Furthermore, the aromaticity index classifications – that is, the percentage of compounds identified as aliphatic, aromatic, or condensed aromatic (AI_{mod} < 0.5, 0.5 - 0.67, > 0.67, respectively),[62] is essentially the same (Figure 5.10b). If anything, the new normally distributed data shows even more aliphatic and fewer condensed aromatic compounds, further emphasising the point of Section 5.1.6. This analysis allowed me to reaffirm the conclusions of the original analysis: ESI and LDI are complementary techniques and both should be used for fuller characterisation of a complex NOM. This conclusion was based entirely on the presence/absence of peaks, and NOT their relative intensities. Therefore, at the most basic level, the fact that the ESI spectrum had a bimodal distribution rather than the typically presented normal distribution is insignificant.

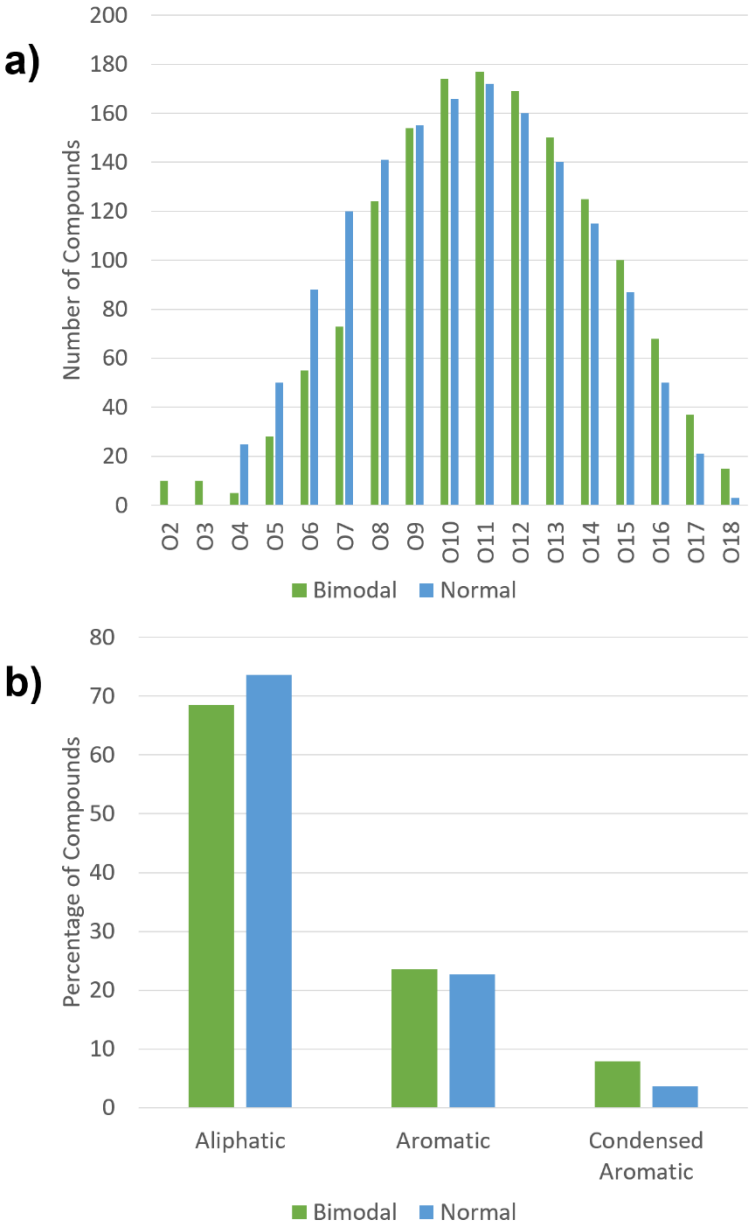


Figure 5.10 a) Heteroatomic class distributions for the bimodal and normally distributed datasets between *m/z* 250-600, b) compound categories (%) according to aromaticity index (AI_{mod}) for the bimodal and normally distributed datasets. For definitions see text.[102] Reprinted with permission from Anal. Chem., 90/9, Kew. W., Blackburn. J.W.T., and Uhrin. D., Response to Comment on "Laser Desorption/Ionization Coupled to FTICR Mass Spectrometry for Studies of Natural Organic Matter, 5968-5971. Copyright (2018) American Chemical Society.

An interesting point to note is that typically presented SRFA ESI spectra[71, 79, 99, 103] are better described as a skew normal distribution with a significant tail extending to 1000 *m/z* (or potentially above) depending on the SNR. A recent 21 Tesla SRFA ESI-(-) FTICR spectrum shows a broad range of peaks from 200 to 1000 *m/z*, with a maximum intensity between 300-500.[99] The data presented in Figure 3 of the paper by Shaw *et al.* has a non-smooth tail which suggests secondary distributions of ions around 620 and 750 *m/z*. In any case, if one looks more closely at the published SRFA spectra,[71, 79, 104] including my work presented here, these are comprised of a number of normal distributions – at each 14 *m/z* units (Figure 5.9c,d), and within each *m/z* (Figure 5.9e,f) in both the bimodial and normally distributed datasets. These distributions build into each other, and – presumably – reflect the inherent chemical stability, prevalence, and ionisability of certain compound classes present in SRFA. Each 14 *m/z* units represents an additional CH_2 and thus extends the molecular size of very similar structures. Arguably, when trying to understand the nature of SRFA compounds as seen by FTICR MS, it is the per-*m/z* distributions that are chemically more interesting and relevant than an overarching distribution across the entire mass range. The latter is much more liable to be biased by instrumental parameters.

5.4.2 Investigation into aggregation and dimerisation

Another concern was that the bimodal distribution obtained was due to dimerisation as a number of signals in the *m/z* 600 – 800 region showed high intensity. To investigate a possible dimerisation, I first compared our bimodal data with the newly acquired normally distributed data. As an example, in Figure 5.11 the *m/z* 701 region in each of the two spectra is shown; both contain the same dominant peaks. This was

typical for the entire high-end *m/z* part of the spectra. Therefore, if the bimodal data reflect aggregation, this is true of the normally distributed data as well. Furthermore, the above discussed per-14 *m/z* periodic distribution pattern continues in the high masses, which suggests no aggregation – dimers would, presumably, repeat at half this interval, per-7 *m/z*.

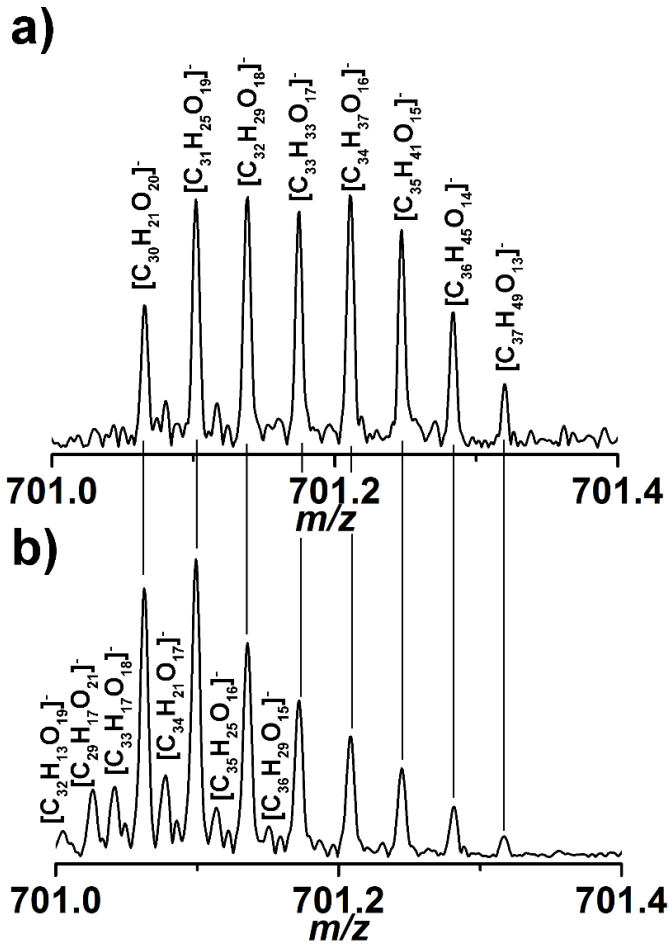


Figure 5.11 A single nominal mass (701 *m/z*) for the normally distributed (a) and bimodal (b) datasets. Assigned ion formula have been annotated. Median error of assignment in this window was 0.72 ppm. Vertical lines have been added to aid comparison.[102] Reprinted with permission from Anal. Chem., 90/9, Kew. W., Blackburn. J.W.T., and Uhrin. D., Response to Comment on "Laser Desorption/Ionization Coupled to FTICR Mass Spectrometry for Studies of Natural Organic Matter, 5968-5971. Copyright (2018) American Chemical Society.

Interestingly, the bimodal data also contain several minor signals, as illustrated within the 701 *m/z* region; these are mostly at the noise level, though some are also evident in the normally distributed data. These peaks could also be assigned molecular formulae, and represent additional important information about the sample. As the number of theoretical combinations of CHO formula increases exponentially at higher masses, having a greater SNR at a higher mass is beneficial and a direct consequence of the instrumental setting that produced the bimodal dataset.

Stenson *et al*. presented positive ion ESI of SRFA with two bimodal distributions; the first between 300 and 1200 *m/z* and the second from 1400 to 2200 *m/z*, approximately.[71] This second distribution is likely a result of aggregation as they used a high concentration of 5.7 mg/mL.[71] Fievre *et al*. also presented apparently aggregated ESI spectra of SRFA; unfortunately, they did not report the concentration of SRFA in their experiment.[70] To the best of my knowledge, no recent paper has presented ESI spectra of aggregated SRFA. Typically, ESI spectra of SRFA are now acquired with low concentrations, < 1 mg/mL.[79] To further address the question of dimerisation, I also attempted to force aggregation. I increased concentration from 0.1 mg/mL as shown in the normal and published bimodal data, to 1 mg/mL, 2.5 mg/mL and finally 5 mg/mL. No significant differences between these low and high concentration FTICR MS spectra was observed (Figure 5.12). At 5 mg/mL, far greater than typically used nowadays, no apparent dimerisation in the mass spectrum was observed. However, at this concentration, spray stability begins to suffer, and I do not believe higher concentration samples would be possible to spray on the used instrument. At these higher concentrations, the IAT had to be severely reduced to 0.01s to prevent overfilling the ICR cell and to continue acquiring high quality spectra. IAT has been

shown to impact on relative peak intensity, with longer IATs allowing for observation of more minor components.[105] Interestingly, the very short IAT presented in[105] had an apparent bimodal distribution for ESI of SRNOM. The bimodal dataset, however, had been acquired with a longer IAT of 1.2s.

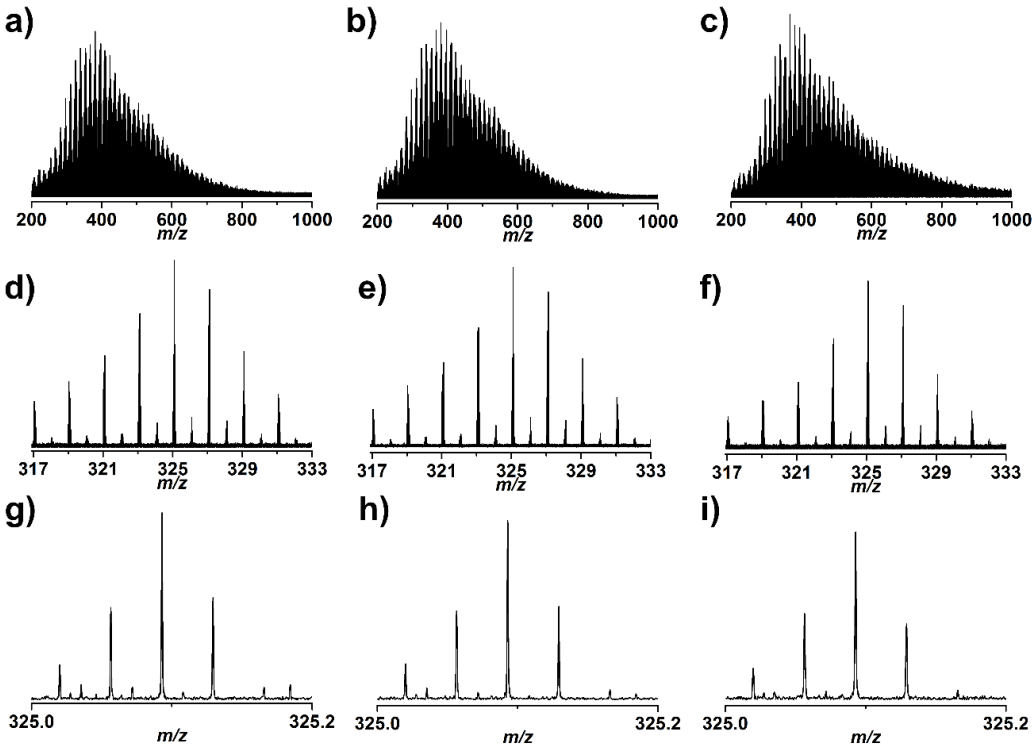


Figure 5.12 ESI FTICR mass spectra for SRFA acquired at different concentrations; left column (a, d, g) at 1 mg/mL, middle column, (b, e, h) at 2.5 mg/mL, and right column (c, f) i) at 5 mg/mL. Top row shows the broadband spectrum, the middle row shows a per-14 *m/z* distribution, and the bottom row shows a single *m/z*.

Furthermore, I experimented with a variety of collision cell voltages for the 2.5 mg/mL sample concentration – including 0 V (not recorded), 0.5 V (typical acquisition voltage), 2 V, 4 V, and 6 V. If the data at 2.5 mg/mL (and thus at all concentrations)

is dimerising, increasing the collision voltage should break up these aggregates before causing further fragmentation. Increasing the collision voltage to 4 V made no substantial difference to the profile, though it did increase the relative intensities of most peaks. At 6 V, the profile of the spectrum shifted marginally to lower masses, from around *m/z* 380 to *m/z* 320, suggesting some degree of fragmentation.

5.4.3 High and low mass species

In my experience, and as discussed elsewhere,[71] the intensity, or abundance, of any given signal in a complex mixture ESI-FTICR mass spectrum is the function of several variables, only one of which is the sample concentration. The signal intensity will also depend upon pH, solvents, instrument condition, and a host of instrument parameters including source parameters and transfer optics parameters.

ESI SRFA spectra are typically presented as a normal distribution between *m/z* 200-700.[105] However, SRFA is expected to contain compounds smaller and larger than this. Orbitrap ESI-MS analysis shows compounds in SRFA between *m/z* 100-600, though Remucal *et al*. acknowledge that non-mass spectrometry approaches report molecular weight ranges for SRFA between *m/z* 800 and 1500.[18]

I would therefore like to suggest that it would be wrong to accept certain presentation of FTICR MS spectra of NOM, or any complex mixture, as standard, or the only appropriate way. To demonstrate this, data was acquired with the spectrometer tuned for low and high mass compounds, and observed (and assigned) compounds as small as *m/z* 109, and as high as *m/z* 1035 were identified.

To prove the usefulness of this approach, I assessed whether the species being ionised are truly a continuation of the previous mass range. To do this the assignments for two ranges, *m/z* 270 – 320 (low mass range) and *m/z* 700 – 750 (high mass range) were compared against the standard spectrum. Starting with the lower mass range, comparing the low and standard spectra assignments, 106 and 137 formulae were identified from each spectrum, respectively. Of the 106 low mass formulae, 102 were also found in the standard spectrum, and manual inspection of the remaining 4 showed they were present in the standard spectrum but fall below the SNR of the peak picking. For the standard to high mass comparison, 156 and 402 formulae were observed from each spectrum, respectively. The significant increase in the number of assigned formulae, for this range, in the high mass spectrum can be attributed to a much higher SNR. From the 156 standard spectrum formulae, 155 were also seen in the high mass spectrum, manual inspection of the remaining one formula's peak showed it hadn't been assigned as it falls just outside the assignment threshold of 500 ppb, however, the peak was still present.

From these comparisons, it can clearly see that the low and high mass spectra are truly a continuation of the standard spectrum and allow formulae to be assigned well past the traditional ranges for ESI of HS samples. The benefits of this windowed acquisition is exemplified on the case of high molecular mass spectrum, where substantially more peaks were observed in the transition region.

Analysing the low *m/z* spectrum (Figure 5.13) it was possible to identify 802 peaks. Of these 477 were assigned to monoisotopic formulae with an average absolute error of assignment of 284 ppb. 45 peaks were assigned formulae containing one ^{13}C with an average absolute error of assignment of 153 ppb. Combining both values represents

an assignment of 65.1 %. The formulae identified contained species from O1 to O9 with a maximum number of formulae identified for O4.

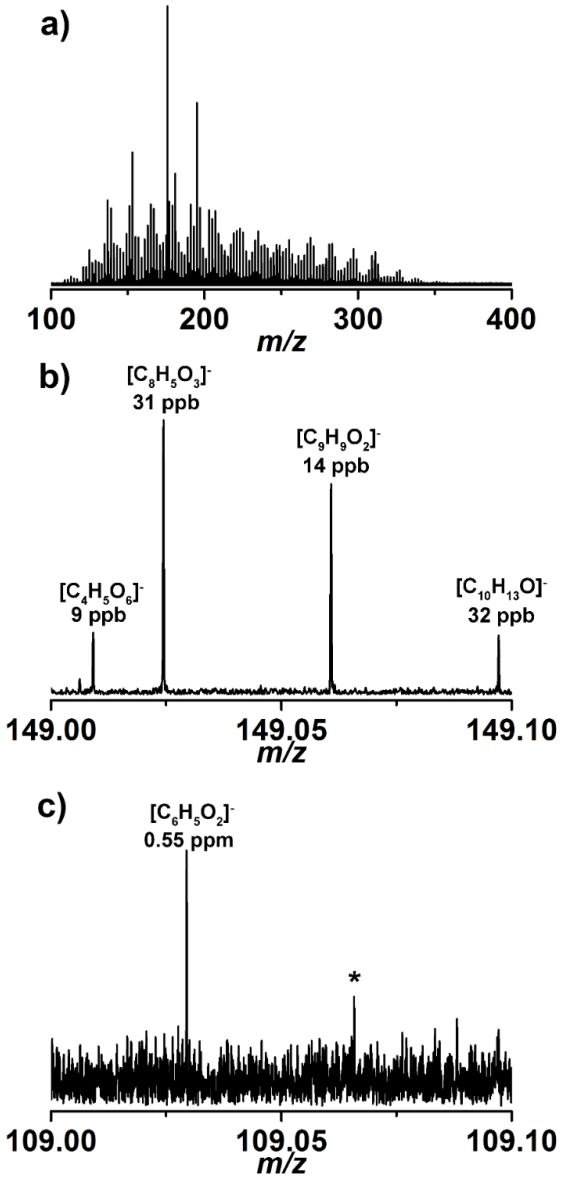


Figure 5.13 ESI-(-) FTICR mass spectra for SRFA when tuned for low masses. a) Shows the broadband acquisition, b) shows a single nominal mass at *m/z* 149, and c) shows the lowest assigned mass observed for SRFA in these experiments at *m/z* 109. The point labelled by an asterisk (*) was not assigned as it was below the peak picking thresholds.

Tuning the spectrometer for the higher mass range I used the quadrupole to filter out any species lower than *m/z* 700. In doing so it increased the SNR of all species above this cut off point. Figure 5.14 show the broadband acquisition of this range, with peaks extending as high as *m/z* 1200. It was possible to calibrate confidently up to *m/*z 1035, at which point the SNR deteriorated. Using a SNR threshold for peak picking of 10 it was possible to identify 3937 peaks. Of these 1922 were assigned to monoisotopic formulae with an average absolute error of assignment of 217 ppb. 900 peaks were assigned formulae containing one carbon-13, with an average absolute error of assignment of 241 ppb. Combined proportion of assigned peaks reached 71.7 %. The identified formulae contained species from O13 classes up to O28 classes with a maximum number of formulae at the O21 class.

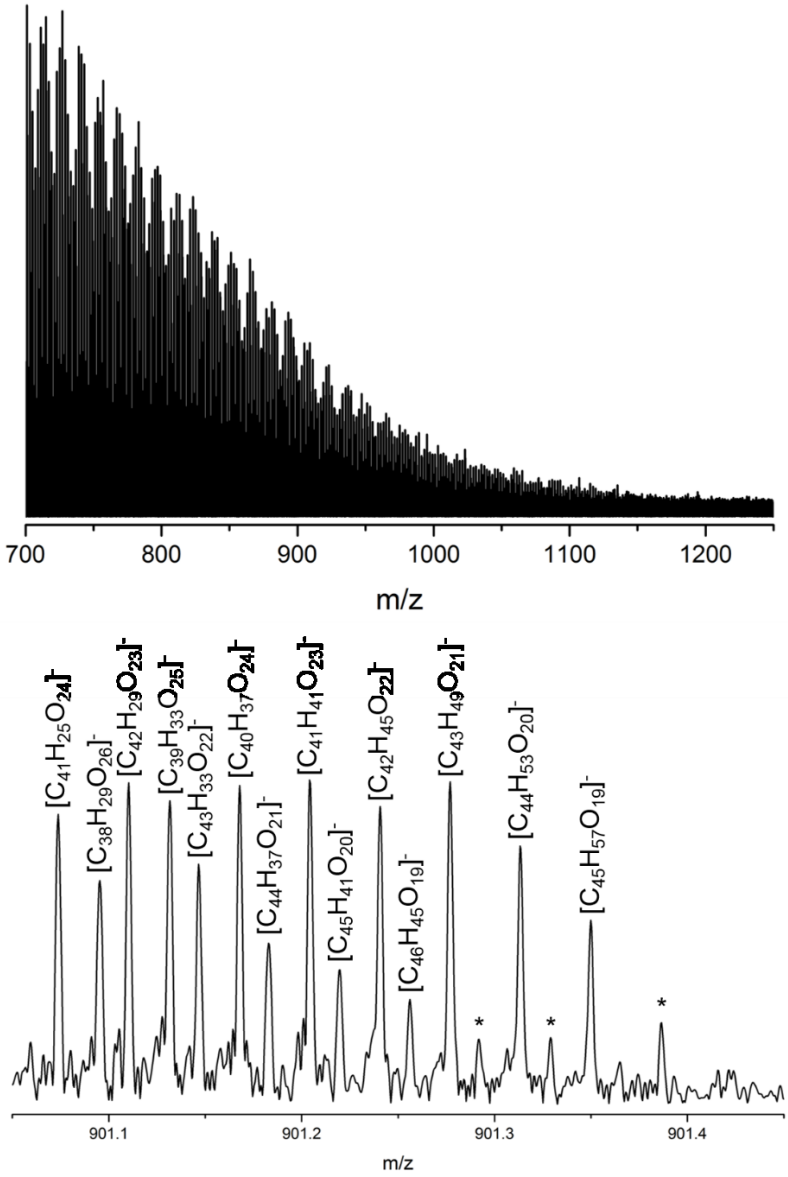


Figure 5.14 Negative ion FTICR mass spectra of SRFA tuned to high mass species. Displayed is broadband acquisition and a single nominal mass at 901 *m/z* labelled with assignments. The peaks labelled by an asterisk (*) were not assigned as they are below the peak picking thresholds.

Van Krevelen diagrams were produced for all three mass ranges (Figure 5.15). Comparing these shows that whilst each spectra ionises a broad range of compound types, the standard spectrum has a little more chemical diversity than the high mass spectrum and the low mass spectrum has much more chemical diversity, with its formulae spreading to a much wider range of H/C and O/C ratios.

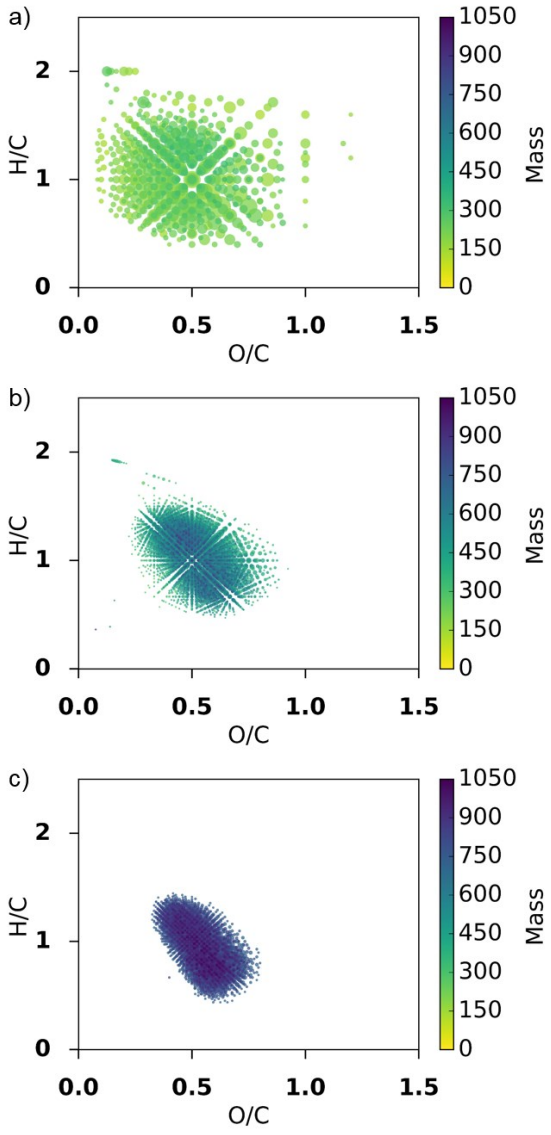


Figure 5.15 van Krevelen diagrams for the low mass spectrum (a), standard spectrum (b) and high mass spectrum (c). The size of the point is relative to the normalised intensity of the peak within the spectrum.

Both low and high mass oxygen species identified fit as a continuation of the oxygen trend seen for the 'standard' mass range (Figure 5.10). This observation of identifiable species seen well above *m/z* 700 agrees with what has also been seen with a magnetic field of 21 Tesla.[99]

For all three mass ranges, $AI_{(mod)}$ was calculated, and by combining and comparing the three, an interesting phenomenon can be seen. In order to best show this the data sets were cut so that the low-mass spectrum (Figure 5.13) described *m/z* 100 – 300, the 'standard' spectrum (Figure 5.9) described *m/z* 300 – 700, and the high-mass spectrum (Figure 5.14) described *m/z* 700 – 1000. Figure 5.16 shows a plot of $AI_{(mod)}$ against mass for all three of the ranges.

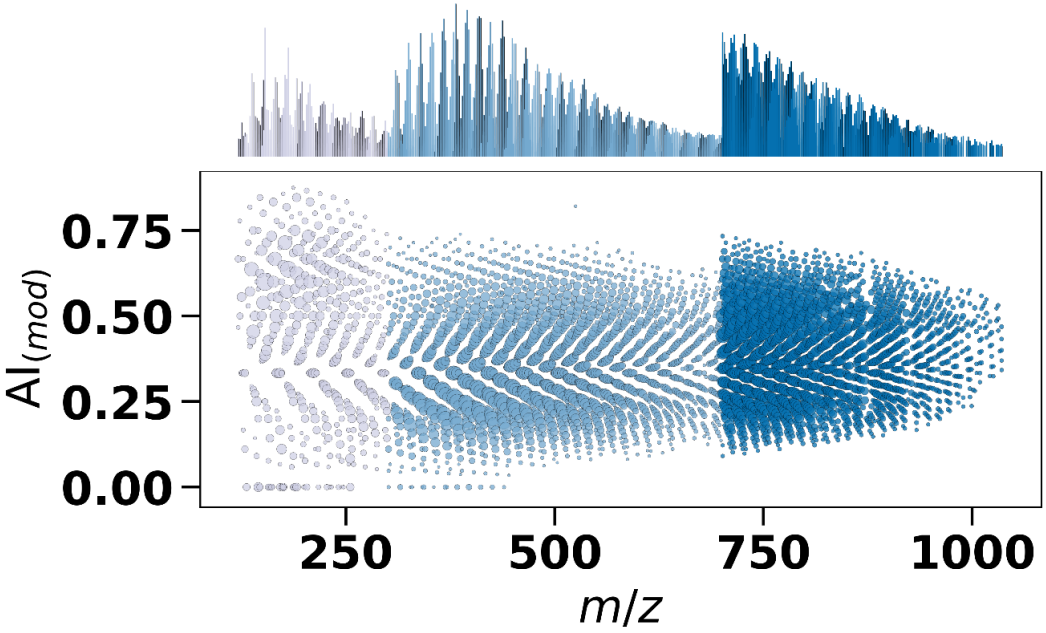


Figure 5.16 $AI_{(mod)}$ against *m/z* of all assigned formulae in SRFA for the three data acquisition methods; low mass *m/z* 100 – 300 (light blue, left); standard *m/z* 300 – 700 (blue, central); high mass *m/z* 700 – 1000 (dark blue, right). The overlay of these three spectra is shown above. Point size represents normalised peak intensity across all three spectra.

A trend can be seen that as the mass increases the total range of $AI_{(mod)}$ values decreases. A deviation from this occurs at 700 *m/z* due to the SNR of peaks in the 'standard' spectrum dropped off, this can be seen from the MS spectra shown at the top of the 2D plot in Figure 5.16. Additionally, this trend is most prominent for the low mass spectrum, where the range of $AI_{(mod)}$ values is at its greatest. This suggests that, of the species identified by ESI-(-), the higher mass species are more chemically similar to one another than the lower ones, in terms of aromaticity. This agrees with the van Krevelen diagrams presented in Figure 5.15. While I cannot say with certainty what causes this I can suggest some possibilities. It could be an instrumental effect in that high mass species that are highly condensed or highly aliphatic are more difficult to produce gas phase ions from, are less ionisable, or are simple in such low concentration that they are outcompeted by other more abundant ions. Additionally it could be sample related, as HS comprises of organic matter that is at various stages of incomplete degradation. This would indicate that the less degraded, high mass, species more closely resemble the parent structures, such as lignin. Whereas the more degraded, low mass, species contains a much more diverse range of chemical structures, both aliphatic and aromatic. From this it is possible to theorise that these large structures contain a mixture of aliphatic, including carbohydrate, and aromatic regions that only begin to chemically diversify as they undergo degradation.

In conclusion, I have demonstrated that the intensities of peaks of FTICR MS spectra of NOM heavily depend on the acquisition parameters and therefore they should be interpreted with caution. Setting the parameters to capture different *m/z* regions is beneficial as it allows more compounds to be detected and thus to obtain a fuller picture of the complexity of NOM.

Chapter 6: Deutero-methylation of humic substances for carboxylic acid functionality analysis by FTICR MS

Due to constantly increasing understanding of the mechanism behind reactions, modern isotopic labelling can be performed with a high level of selectivity. Kühnel *et al.*[91] discerned the mechanism of methylation from using TMS-diazomethane, a safer substitute for diazomethane. They found that an intermediate state of the reaction required a pseudo-ring state that was only possible when the double bonded oxygen atom of a carboxyl group was present. This resulted in the reaction being highly selective to methylation at carboxyl sites; this can be seen in step 8 of Figure 6.1.

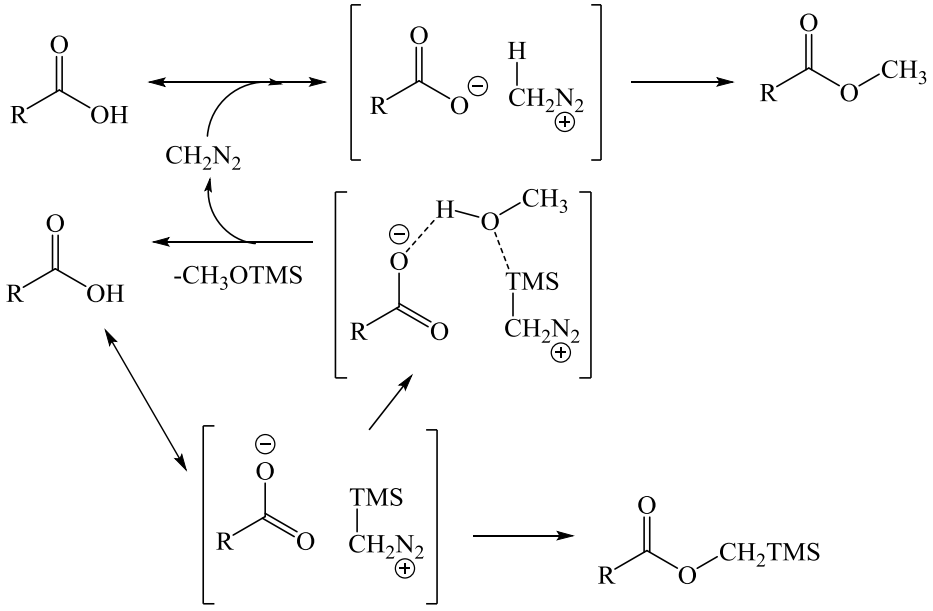


Figure 6.1 Reaction scheme for carboxylic selective methylation using TMS-diazomethane as present by Kühnel *et al.*[91]

Additionally, they demonstrated that by replacing the methanol solvent with methanol-d_1, CD_3 methyl groups would be added instead of CH_3. D has a nominal mass of 2 Da but an exact mass of 2.014102 Da whereas H_2 also has a nominal mass of 2 Da but an exact mass of 2.01565 Da. The difference thus is 1.548 mDa and whilst that is a small difference, it can still be resolved using a very high-resolution mass spectrometer, a FTICR. In order to ascertain the reliability of this reaction it was tested on model compounds before being used to label HS.

6.1 Initial reaction and model compound testing

The reaction was tested using model compound **1**. All glassware was additionally washed with MeOH for the CH_3 methylation and MeOD for the CD_3 methylation, prior to use. 35.3 mg of **1** was dissolved in 6.5 mL solvent solution (20 % MeOH, 80 % toluene). The solution was stirred for 5 hours before being transferred to an inert N_2 environment. 0.13 mL TMS-diazomethane added followed by a further 30 minutes stirring. TMS-diazomethane adds a yellow/green colour to the solution, as the reaction progresses, N_2 is evolves and the yellow/green colour of the solution gradually fades. After 30 minutes the reaction is quenched by addition of 2-3 drop of concentrate AcOH (until the solution is colourless). The solution was evaporated to dryness to give a colourless oil (40.1 mg) which was then dissolved in $CDCl_3$ for NMR analysis. While the reaction has been shown to reliably methylate with close to 100% yield, a complete reaction would only produce 38.9 mg of product. This suggests that one or multiple of; contaminants; incomplete drying; or by-products are present to some degree.

The NMR spectra showed successful methylation with a small amount of starting material being present. A peak around 0 ppm, was observed that could be either a

contaminant or by-product that would explain the additional yield. Additional experimental improvements considered at this stage were the reduction of the initial mixing time (from 5 h to 1 h) and the choice of the solvent system to be optimised for HS solubilisation; either through varying the percentage of MeOH used or by changing toluene for a different solvent.

6.2 By-product identification

The next reaction performed was identical as described in Section 6.1, except for the reduction in the initial mixing time from 5h to 1h. 34.5 mg of **1** was dissolved again. NMR analysis confirmed the reaction worked similarly to the previous experiment, however, smaller peaks at 0.28, 0.18, 0.14 and 0.08 ppm could be seen. A Si INEPT experiment was performed and two major peaks were identified at 0 and -22 ppm, these can be seen on the vertical trace of Figure 6.2. A $^{1}H,^{29}Si$-HMBC spectrum (Figure 6.2) revealed multiple cross peaks correlating with these ^{29}Si signals.

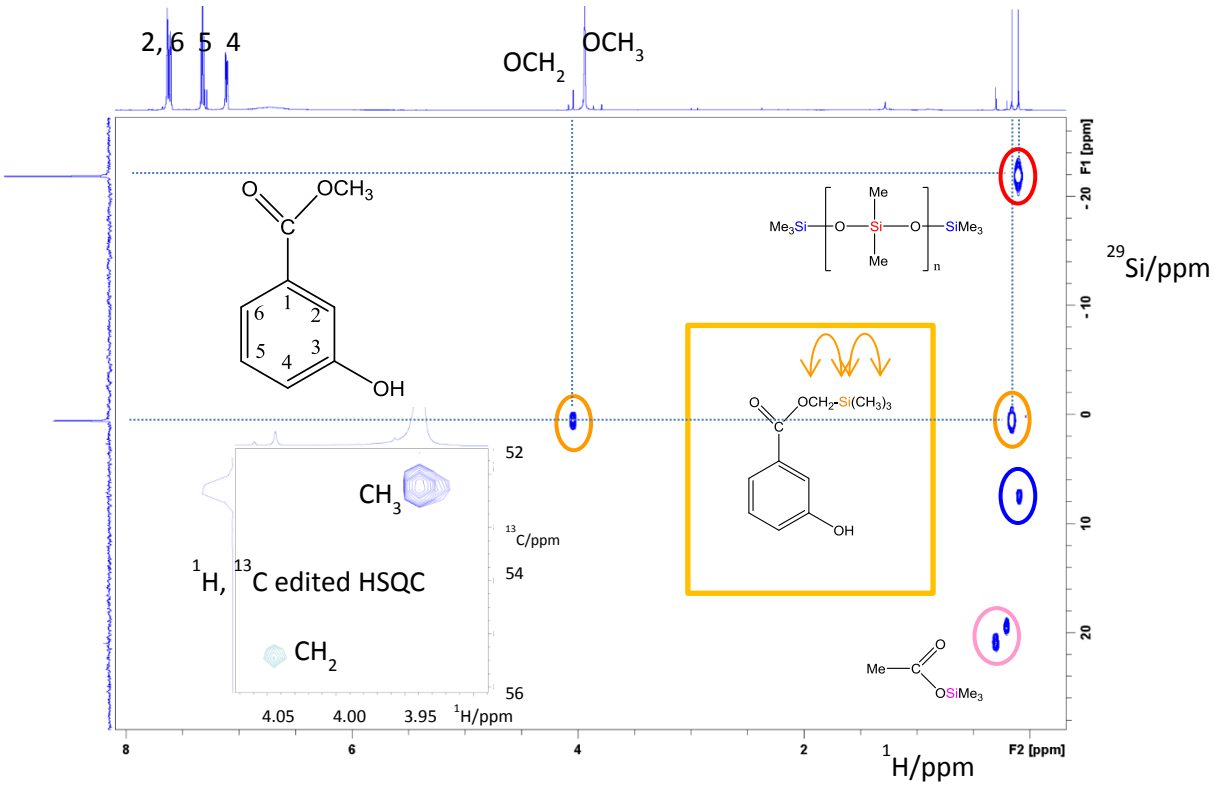


Figure 6.2 ^{1}H, ^{29}Si-HMBC spectrum of the reaction mixture described in the text. Inset: ^{1}H, ^{13}C edited HSQC zoomed onto the methoxy region, which also contains a negative crosspeak belonging to an O-CH_2 group.

Six cross peaks can be identified, five of which are around 0 ppm on the ^{1}H axis and one of which is in the 4 ppm region. This isolated cross peak at 4.05 ppm can only be ascribed to the silylated by-product, where instead of the methylation incorporating a CH_3 group it has instead added a $^{-}CH_2$-TMS group; the Si of this group resonates at 0.16 ppm. The crosspeak assigned to this group appeared negative, in the ^{1}H, ^{13}C HSQC spectrum, indicating that this indeed is a CH_2 group. Of the remaining 4 cross peaks two are produced by a single major compound (0.1 [^{1}H]/ 7 and -22 [^{32}Si]) and two are from separate minor compounds. This major compound is likely to be silicone

grease that has contaminated the system at the glass joints on the gas line during both reaction and evaporation steps. One of the minor compounds (0.3 ppm [^{1}H]/ 21 ppm [^{32}Si]) was attributed to AcO-TMS, produced when the reaction is quenched, and the other (0.21 ppm [^{1}H]/ 19.5 [^{32}Si]) to MeO-TMS, which is formed during the intermediate step of the reaction.

Fundamentally these minor compounds and silicone grease aren't of huge concern as they will either not be ionisable and thus not detectable by FTICR MS or are significantly different to any HS molecules and will therefore be easily identifiable. The by-product of the starting material however is of more concern as it impacts the yield of the reaction. With this in mind I decided to reduce the amount of TMS-diazomethane used from 0.13 mL to 0.05 mL, in order to reduce the large excess of reagent down to 1.2 molar equivalents. This should reduce the amount of by-products formed.

6.3 Effects of MeOH:toluene ratio on yield

The reaction was repeated with the previously stated change to reduce TMS-diazomethane to 0.05 mL. 35.2 mg of reagent **1** was used to produce 44.1 mg yield and the concentration of the by-product peaks was greatly reduced. However, when using reagent **1**, the final reaction vessel contained 2 substances, a colourless oil which contained the product and was soluble in $CDCl_3$, and a white powder which was insoluble in $CDCl_3$ and assumed to be remaining starting material. Taking this white powder and dissolving it in DMSO-d_6 followed by analysis using ^{1}H NMR experiments confirmed this assumption.

The reaction was repeated, however, the entire product was dissolved in anhydrous-DMSO-d_6 in order to compare the amounts of starting material and product. Using ^{1}H and $^{1}H,^{13}C$-HSQC experiments the peaks for the ring protons were examined for a single proton where the starting material and product were resolved from one another. This separation can be seen for H4 of both the product and starting material (Figure 6.3). Using this I was able to compare the relative intensities of these two peaks in order to establish the yield relative to the total reclaimed material. This was calculated to be 60 %. Despite the reduction of the TMS-diazomethane volume added having negatively affected the yield it also provided a sudden opportunity, due to the yield no longer being effectively 100% we are in a position to alter the solvent system and assess its effect on the yield, whether it is positive, negative or ineffective.

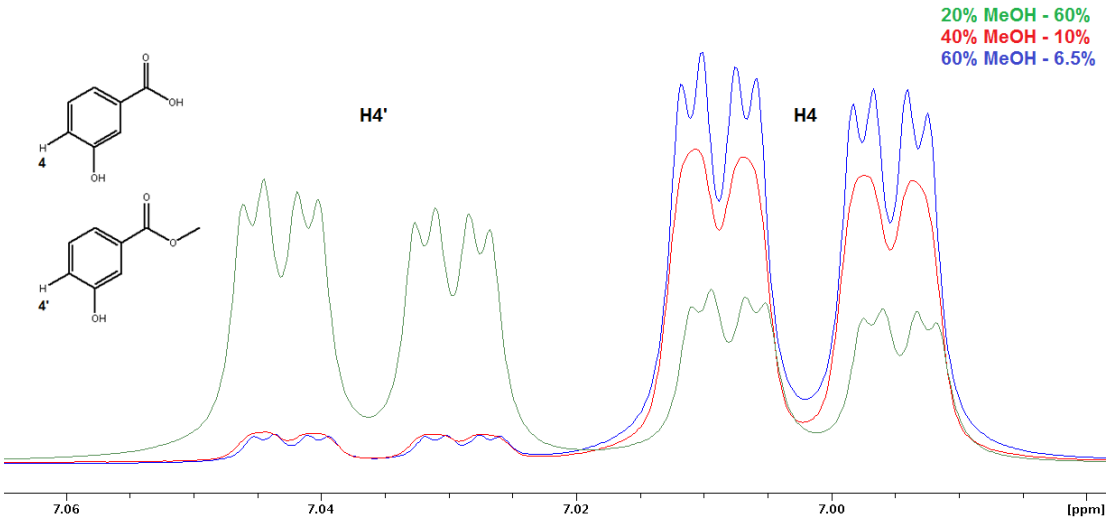


Figure 6.3 NMR spectra zoomed into the H4 region of the starting material and product peaks, showing the increased yield when using 20 % MeOH (green) over 40 % (red) and 60 % (blue).

HS solubility is a major concern when performing reaction in organic systems as it may not be soluble in non-polar organic solvent, such as toluene. MeOH was successfully used to dissolve HS in previous work (sections 3 and 4), therefore I decided to repeat the reaction with increased ratios of MeOH to toluene. This increase in MeOH volume would make the solvent solution more suitable for solubilising HS. Two additional solvent systems were tested, toluene with 40 % and 60% MeOH. The reactions were performed exactly as above with the only difference being the solvent concentration and an adjustment in the time to evaporate, as the higher methanol volume resulted in faster evaporation times. The H4 proton intensities were compared and the yields were found to be 10 % and 6.5% for the 40 % MeOH and 60% MeOH reactions, respectively. An overlay of the ^{1}H NMR spectrum for the reaction mixture produced by the three reactions can be seen in Figure 6.3, zoomed into the H4 region. This visually shows the difference in effectiveness for the three reactions. Ultimately this showed that 20 % was the most effective MeOH volume, which means that the solubility may be limiting the performance of the reaction on HS.

6.4 Methylation and deutero-methylation of a model mixture

This was the last standard reaction to run on model compounds before beginning to test the deuteration reaction. I decided to perform the reaction on a mixture of models compounds to establish whether I obtain maximum yields for all starting materials using the original TMS-diazomethane concentration. A mixture containing roughly 0.04 mmol of several model compounds; 5.9 mg **1**; 6.1 mg **2**; 6.6 mg **3**; 6.3 mg **4**; 6.5 mg **5**; 7.2 mg **7** was reacted with the original volume of 0.15 mL TMS-diazomethane as described in Section 3.2.1. As expected no starting product was observed for any of

the models and all 6 products were identified by NMR. This is shown alongside the assignments in Figure 6.4.

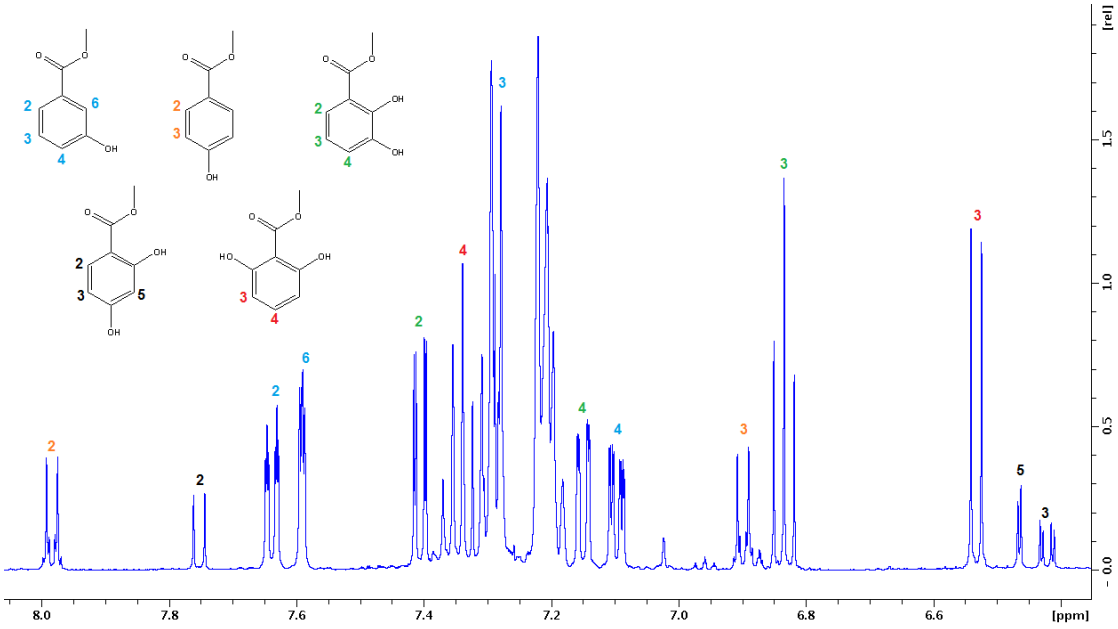


Figure 6.4 Aromatic region of a ^{1}H NMR spectrum for the model mixture with resonance assignments.

The same models were used in order to test the efficiency of deuteration. A number of additional steps were included to improve this efficiency. All glassware was prewashed with MeOD. The starting materials; 6.1 mg **1**; 6.1 mg **2**; 6.6 mg **3**; 6.3 mg **4**; 6.5 mg **5**; 7.2 mg **7** had their labile protons pre-exchanged with deuterium by twice washing with excess MeOD followed by evaporation to dryness. Finally 0.15 mL TMS-diazomethane was stored under N_2 in 0.3 mL MeOD overnight. Visual observation of the TMS-diazomethane the following morning revealed a significant colour loss to nearly colourless. Upon addition to the reaction vessel, no N_2 could be seen evolving from the reaction. This suggests that the TMS-diazomethane had

degraded overnight. 0.15 mL fresh TMS-diazomethane was added and N_2 bubbles began forming instantly, this confirms that the exchanged reagent had no activity and thus no N_2 bubbles were seen. Each starting reagent only contained one carboxyl group, resulting in only a single site for methylation. Figure 6.5 shows a modified reaction scheme as shown in Figure 6.1 with the sources of deuterium highlighted. From this it can be see that each individual deuterium atom is incorporated separately within the mechanism. Assuming this incorporation doesn't occur at 100 % efficiency that leaves the potential for four products per starting material. These four are methylations that add a CH_3, CH_2D, CHD_2 and CD_3 group.

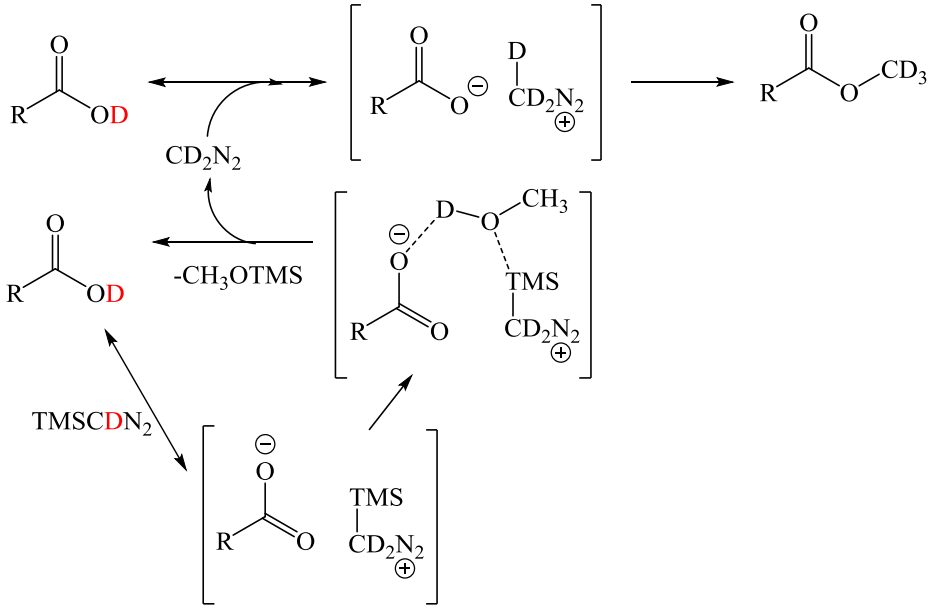


Figure 6.5 Modified reaction scheme for deutero-methylation showing the three exchange locations of H for D (highlighted in red) required to produce a fully deuterated, CD_3, methyl ester.

In order to maximise the concentration of any single product in the final sample and increase the likelihood of detection by experiments on real HS only the CD_3 product is desirable. NMR was used in order to identify the presence of these potential 24 products.

Initially two ^{1}H, ^{13}C HSQC spectra of the product mixture were acquired (Figure 6.6), the first of which had both hydrogen and deuterium decoupling, and the second which had only hydrogen decoupling. These enabled the identification of incomplete deuteration by-products within the mixture. With the deuterium coupled spectrum showing coupling for the D_1 and D_2 products. However as the D_3 product contains no methoxy protons, no crosspeak can therefore be seen to confirm its presence.

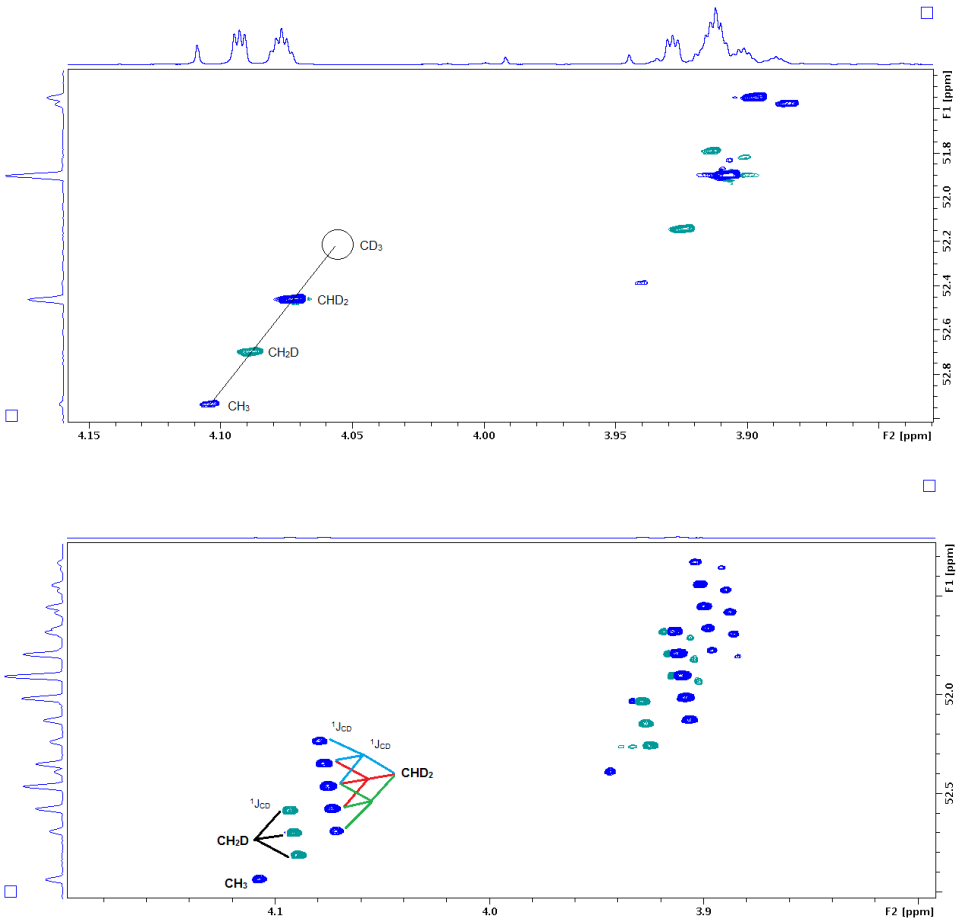


Figure 6.6 ^{1}H, ^{13}C HSQC spectra of the product model mixture. Top is ^{1}H and ^{2}D decoupled and shows assignments along with theoretical position of CD_3 although it cannot be identified by the experiment. Bottom is ^{1}H decoupled and confirmed the assignment of the deuterium containing by-products. The annotated peaks on the left belong to one model compounds products that are separated from the rest which can be seen on the right.

Three ^{13}C spectra of the product mixture were acquired, one with deuterium decoupling, one with both hydrogen and deuterium decoupling and one with hydrogen and deuterium decoupling as well as DEPT editing. Focussing on the methyl ester

region (~52 ppm), as shown in figure 6.7, the combination of these three spectra can be used to identify the peaks in this region.

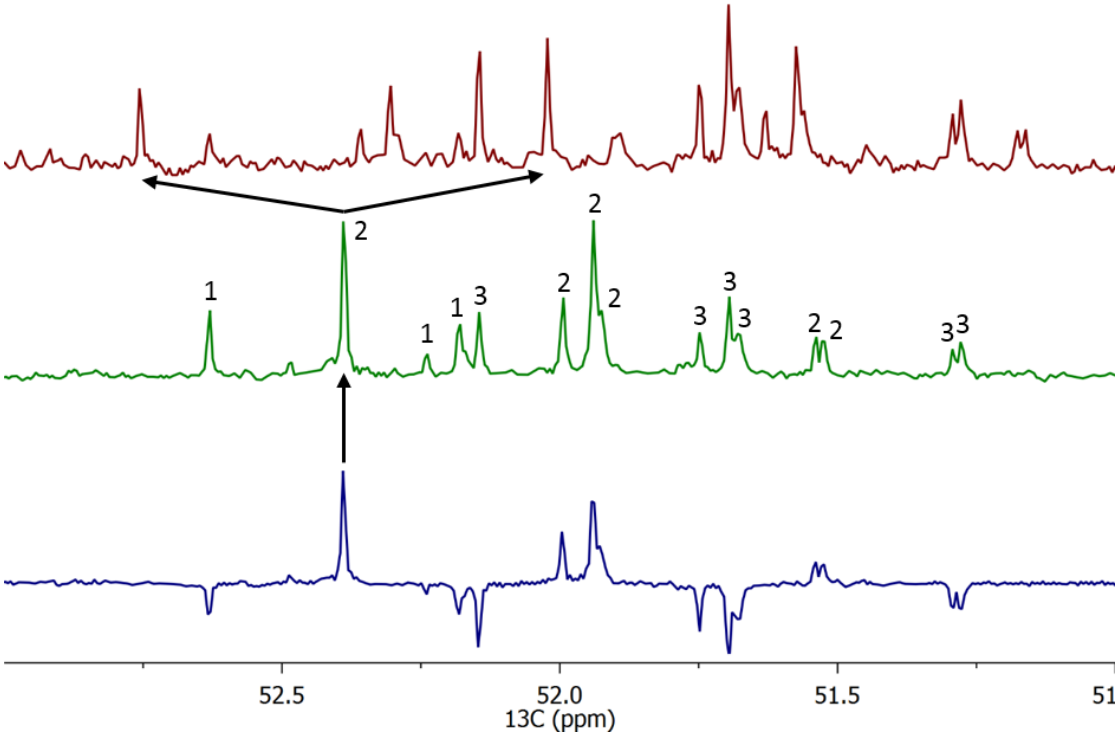


Figure 6.7 ^{13}C NMR spectrum focussed the methyl ester region for the methylated model mixture. The spectra are (top to bottom) ^{1}H decoupled, ^{1}H and ^{2}D decoupled, ^{1}H and ^{2}D decoupled as well as DEPT edited. Three types of methyl ester can be identified; CD_3 denoted by (3); CHD_2 denoted by (2) and CH_2D denoted by (1). CH_3 compounds are presumed to be too low concentration for detection.

The DEPT editing causes all CH_2 and quaternary carbons to point down and any CH_3 and CH carbons to point up. As this spectrum is both hydrogen and deuterium decoupled all peaks will be singlets. Comparing these peaks and the middle (hydrogen and deuterium without DEPT) spectrums peaks to the top spectrum identifies which

groups contain hydrogen. Any groups that contain hydrogen (CH_3, CHD_2 and CH_2D) will cause the peaks to be split, shown by the arrows on Figure 6.4, and the peaks will shift. Peaks belonging to CD_3 groups will not be split by hydrogen and so will remain unchanged across all three spectra. This means it is possible to assign any peaks associated to CD_3 and predict which of the remaining peaks belong to the other forms of by-product by combining the DEPT and coupled spectra. Using this information alongside Figure 6.4 allowed me to identify three CH_2D species at very low intensity, six CHD_2 species at high intensity and six CD_3 at moderate intensity. While it is promising that all the desired CD_3 products were identified, they were not at sufficient intensity to attempt on a real HS sample due to the large amount of by-products formed. CHD_2 was the major product and as the pre-exchange of TMS-diazomethane failed this is the likely cause of the final deuterium being missing. As Figure 6.4 shows, this pre-exchange is where one of the deuterium is incorporated. Therefore it is important to develop a way to perform this pre-exchange without it resulting in the degradation of the TMS-diazomethane.

6.5 Improving pre-exchange of TMS-diazomethane

In order to establish if the previous pre-exchange tested was degrading the reagent the reaction was performed on a single model compound, 35 mg **2**. 0.15 mL of TMS-diazomethane was stored under N_2 in 0.3 mL of MeOD overnight. There was no initial colour loss in the first hour, however, by the following morning there was significant colour loss. Again, no N_2 was formed upon the addition of TMS-diazomethane and only **2** could be identified by NMR. This confirms that the degradation did occur overnight, only explanation is that air could be leaking into the vessel over a longer

period of time, and only a small amount would cause the gradual degradation. In order to tackle this, I decided to attempt the pre-exchange over a much shorter time period. 0.15 mL TMS-diazomethane was stored in 0.5 mL MeOD under N_2 for 30 minutes before addition to the reaction vessel with **2**. Initially no N_2 was formed (which is usually instantaneous upon addition) however ~10 s after mixing N_2 began forming. It was theorised that the cause was likely to be one of two reasons: either the degradation had already begun but due to the shorter exchange time had not reached completion, or that the reaction progressed slower when TMS-diazomethane contained a deuterium instead of hydrogen. The simple way to answer this was to examine the product by NMR and determine how much starting material was present. From the ester region of the 1D ^{13}C spectrum two products could be identified, the CHD_2 and CD_3 products, which presented as a quintet and a heptet respectively. The intensities of these peaks suggest a similar ratio of the two products however this is not reliable as it is a comparison of a tertiary and a quaternary carbon, this is due to the NOE enhancement of the tertiary carbon signal from the hydrogen. The ^1H spectrum is much more reliable for quantification and comparison of the product aromatic peaks with the by-products methyl ester peaks showed that the CHD_2 product accounted for 9 % of the total methylated product. Similarly to previously discussed, the intensities of fully separated aromatic protons in the ^1H spectrum were used to quantify the ratio of product to starting material. The final calculations resulted in 20 % of reactant **2**, 7.2 % CHD_2 product and 72.8 % CD_3 product, structures given in Figure 6.9.

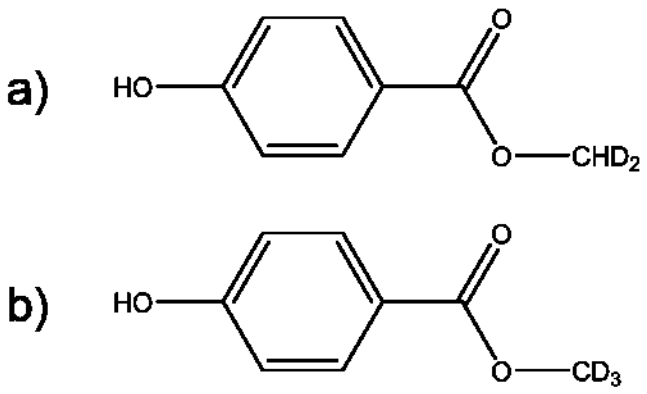


Figure 6.9 Structures of reaction showing a) CHD_2 and b) CD_3 methylation products.

The increase in the amount of triply deuterated product suggests that the pre-exchange step was causing the large amount of CHD_2 by-product seen in the previous experiment, however even within 30 minutes enough of the TMS-diazomethane had degraded so that 20 % of starting material **2** remained at the end. It was ultimately concluded that this pre-exchange step resulted in a trade-off, a loss of reaction efficiency and yield and in return an improvement in deuteration efficiency. As the nature of HS is very complex, discounting any other factors, there would still be species with no carboxyl groups present in the sample across that would be identified by FTICR MS. Thus the peaks would still be present in the spectra alongside the new methylated product peaks, further increasing the complexity of the sample. With this in mind I decided that, providing that the yield was high enough, the most important factor was a high ratio of CD_3 products to by-products, and the current method had successfully achieved this. Taking this into account, the final step before attempting this reaction on a HS sample was to confirm the ability to successfully ionise and identify these methyl-deuterated using FTICR MS.

6.6 FTICR MS analysis of a deutero-methylated model mixture

A model mixture containing; 6.2 mg **1**; 6.1 mg **2**; 6.6 mg **4**; 6.6 mg **5**; 7.0 mg **6**; 6.6 mg **7**; was prepared and the reaction was performed as described in Section 3.2.2. Following evaporation of the organic solvents the products were dissolved in LC-MS grade H_2O and then lyophilised, this was repeated three times. As the starting materials are pre-exchanged and reacted in MeOD all labile protons, including alcoholic and phenolic protons, will be exchanged for deuterium as well. As the NMR analysis focussed on the methyl ester groups and therefore the hydrogen - deuterium exchange did not affect the results, however, for FTICR MS, these OH to OD exchanges could cause many problems for assignment. Potential back exchanging to protons within the source could make the states of these groups unpredictable. Thus, the dissolve and lyophilisation steps are required in order to revert all OD groups back to OH functionalities. 37 mg product was collected after lyophilisation, this was used to create a 1 mg/mL product solution in 50 % $MeOH_{(aq)}$ which was analysed by FTICR MS.

For all models the major species identified were the starting material and the CD_3 product, as expected. Minor compounds identified included CHD_2 and CH_2D products, ^{13}C isotope peaks, and singly exchanged OD compounds for both the starting materials and products. The major peaks from the spectrum along with marked peaks relating to a single compound, **6** and its products can be seen in Figure 6.10 and Table 6.1. The y-axis in Figure 6.10 is not relative to all peaks as some would not be visible if they were shown to scale, thus the intensities of the assigned peaks are given in Table 6.1. Additionally, all peaks were assigned however I have only focussed and highlighted

model **6** as it had the highest intensities, and therefore was the optimal choice to explore any by-products and their relative intensities.

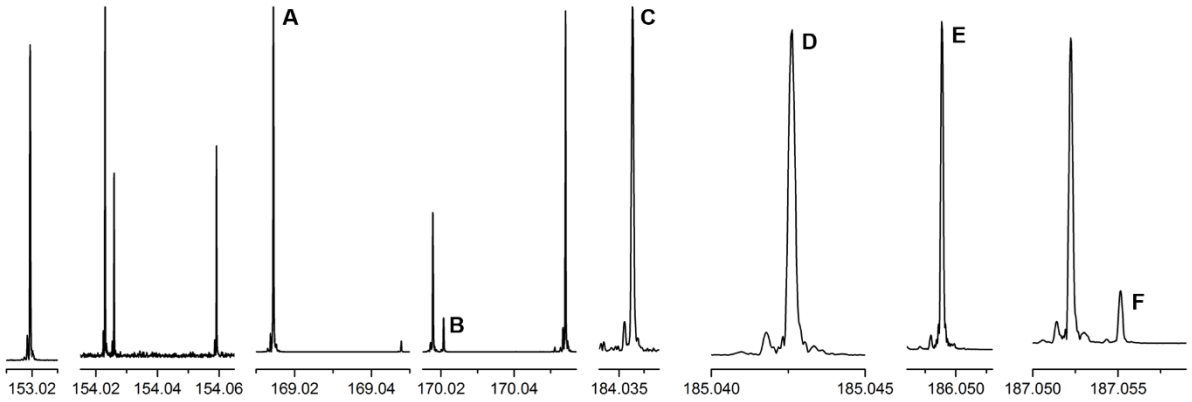
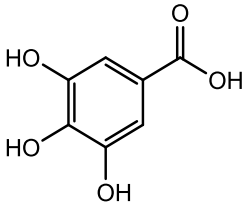
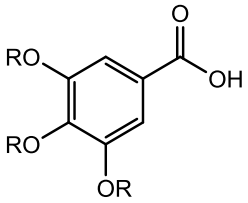
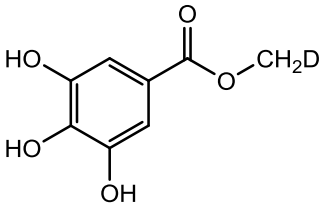
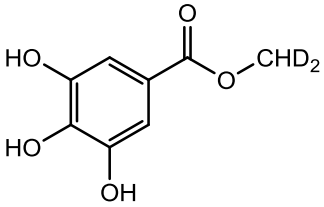
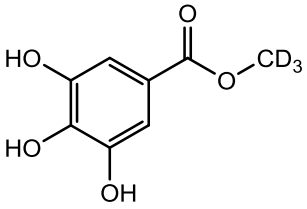
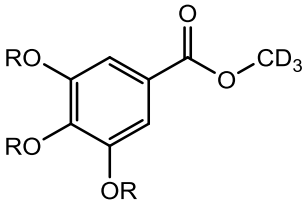


Figure 6.10 FTICR mass spectrum of the methylated model mixture, assignments of all products formed for **6** (A) are labelled (B-F) along with additional data in Table 6.1. y-axis scaled up to match E, exact intensities given in Table 6.1.

Table 6.1 Spectrum assignments of known structures to data highlighted in Figure 6.10

Label	Experimental Mass (Da)	Theoretical Mass (Da)	Products	Intensity ($x10^{-7}$)
A	169.01449	169.01425		226
B	170.02075	170.02053	 R = 2x H, 1x D	3
C	184.03633	184.03618		2
D	185.04263	185.04245		46
E	186.04886	186.04873		335
F	187.05516	187.05501	 R = 2x H, 1x D	5

The intensity ratio of the starting material to the CD_3 product was much more similar than the 20:80 split seen by NMR for the same product. This is a result of MS not being truly quantitative, as the peak intensities are affected by molecular ionisability as well as sample concentration. The starting materials all still have a carboxyl group, which the product doesn't. Therefore, the ionisability difference accounts for the distorted ratio of the compounds (40:60). The OD containing compounds were present, but all at a level lower than even the ^{13}C peaks, suggesting that the lyophilisation step successful removed most of the labile deuterium from the products. The minor peaks caused by the CH_2D and CHD_2 products are the main concern. The ultimate desire for the reaction is to be able to use it to count carboxyl groups, by counting the number of deuteriums incorporated into the molecule where three deuteriums signify one carboxylic acid in the parent compound. The D_1 and D_2 by-products risk producing false positives when combined together. For example, a molecule containing two carboxylic acid groups should contain six deuterons post reaction, however if this molecule reacts to form a CH_2D and a CHD_2 methyl esters, it will only have three deuterons and will appear as if containing only one carboxylic acid. Other problem combinations with the highest possibility are 3 x D_2 and 3 x D_1.

In order to assess the likelihood of these occurrences we compare the intensity of these peaks for the products of **6** as it contains three phenol groups and thus has the highest ionisability. From the total deuteron-methylated product formed, the percentage of CH_2D, CHD_2 and CD_3 were respectively, 0.5 %, 12.1 % and 87.4 %. Considering the two combinations with the highest probabilities (D_1 + D_2 and 3 x D_2) compared to their fully deuterated counterpart (2 x D_3 and 3 x D_3) we can estimate the likely interference of these by-products. The first combination would give a peak intensity of 0.06 % vs

76.4 % and the second gives 0.2 % vs 66.8 %. Therefore, the CD_3 products should theoretically be easily identifiable and whereas any false positive combinations of by-products should be lost in the noise. Considering everything aforementioned we felt we had a suitable grasp on the chemistry of the reaction, both for methylation and deuterium incorporation and could attempt it on a real HS sample.

6.7 Methylation and deutero-methylation of SRFA

In order to test the reaction, a HS standard was selected to undergo methylation. SRFA was chosen due to its greater solubility compared to other HS samples/standards making it more likely to yield full methylation. In order to interpret the results on complex mixture samples it was planned to produce two samples, one that had been methylated with CH_3 groups and another with CD_3 groups. Including deuterium into the formula calculator for assignment greatly reduces the gap between potential formulae, the smallest of which is 0.092 mDa. This represents a difference of 230 ppb at 400 *m/z*, which is smaller than our assignment threshold of 500 ppb and is similar to our average assignment error (see section 4), thus the automated formula assignment will not work for deuterium containing spectra. Due to not being able to use automated assignment, a comparison of the deuterated and original samples prepared with identical conditions was required. A sample methylated with CH_3 groups was used as a starting point to search the sample containing CD_3 species. Upon identification, this could be used to confirm the formula of the parent compound. While this could also be done using the unmodified and deutero-methylated sample, using two methylated sample benefit from a smaller different between the peaks (3 Da per carboxylic acid compared to 17 Da), reducing the risk of incorrect assignment caused by calibration

drift. This also meant that both samples were prepared using identical conditions and should therefore contain identical compounds, albeit in different isotopomers.

The two reactions were carried out using 10 mg SRFA and reduced total volumes of 1.6 mL toluene and 0.4 mL MeOH/D. These reductions were made due to SRFA being a precious sample and FTICR MS requiring much smaller amount of products compared to NMR due to its superior sensitivity. Procedures for CH_3 and CD_3 methylation of SRFA given in sections 2.2.3 and 2.2.4, respectively. In order to maintain identical conditions; the initial pre-exchange and post lyophilisation steps were performed on the undeuterated reaction using MeOH instead of MeOD. N_2 was seen evolving as the reaction progress and there were no notable changes or observations compared to previous reactions on model compounds.

6.8 Acquisition of FTICR mass spectra for methylated and deutero-methylated SRFA

Five samples were prepared, a SRFA standard at 0.1 mg/mL in 50 % $MeOH_{(aq)}$, the methylated product (SRFA-CH_3) at 0.1 mg/mL and 0.5 mg/mL in 100 % MeOH, and the deutero-methylated product (SRFA-CD_3) at 0.1 mg/mL and 0.5 mg/mL in 100 % MeOH, all of which were spiked with 10 % low concentration ESI tuning mix v/v. Pure MeOH was used to dissolve the SRFA-CH_3 and SRFA-CD_3 samples due to the removal of all their carboxyl groups heavily reducing their solubility in water. The spectra were acquired using the instrumental parameters given in Section 3.5.5.

While the SRFA standard could be calibrated and assigned using previously and confidently assigned formulae, these calibration lists were not effective on the SRFA-

CD_3 spectra due to the introduction of deuterons, thus the tuning mix was selected as a suitable starting point to calibrate this potentially difficult dataset. However, in both the SRFA-CH_3 and SRFA-CD_3 samples very little signal belonging to the sample could be seen besides a large contaminant peak at 615.99 *m/z*. This was true of both the low and high concentration samples. Upon investigating this peak, we discovered that it was due to using the tuning mix with 100 % MeOH. The calibrant, as shown in Figure 6.11, binds OH^- to give a peak at 601.99 *m/z*, however with no water present in combination with an abundance of highly ionisable MeOH it was instead binding OMe^-. This resulted in the high intensity contaminant peak that was causing signal suppression of all sample peaks in the spectrum.

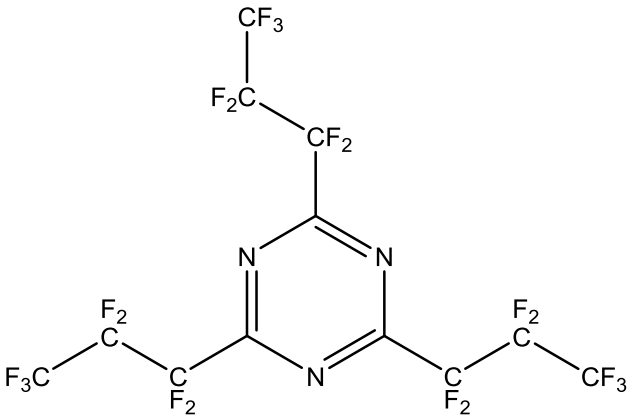


Figure 6.11 Chemical structure of 2,4,6-Tris(heptafluoropropyl)-1,3,5-triazine, one of the calibrants in the used tuning mixture to produce a peak at 601.99 *m/z* upon binding ^-OH.

With tuning mix ruled out as a calibration possibility we were forced to try a different approach. One potential approach was to run the SRFA standard sample and SRFA-CD_3 consecutively, to calibrate the SRFA spectrum in the usual manner and then apply

its calibration as an external calibration onto the SRFA-CD_3 spectrum. The three samples at 0.1 mg/mL were remade and rerun on the instrument as previously stated but without spiking in the tuning mix. Removal of the tuning mix produced familiar spectra of a complex mixture (Figure 6.12). However, there were still a number of large peaks spread across the spectrum at a semi regular spacing. While these peaks weren't so large that nothing else could be seen, as was the case for the previous attempt, they were still large enough to cause a significant amount of signal suppression to the peaks of interest, which adversely affected their intensity and thus their ability to be resolved from the noise.

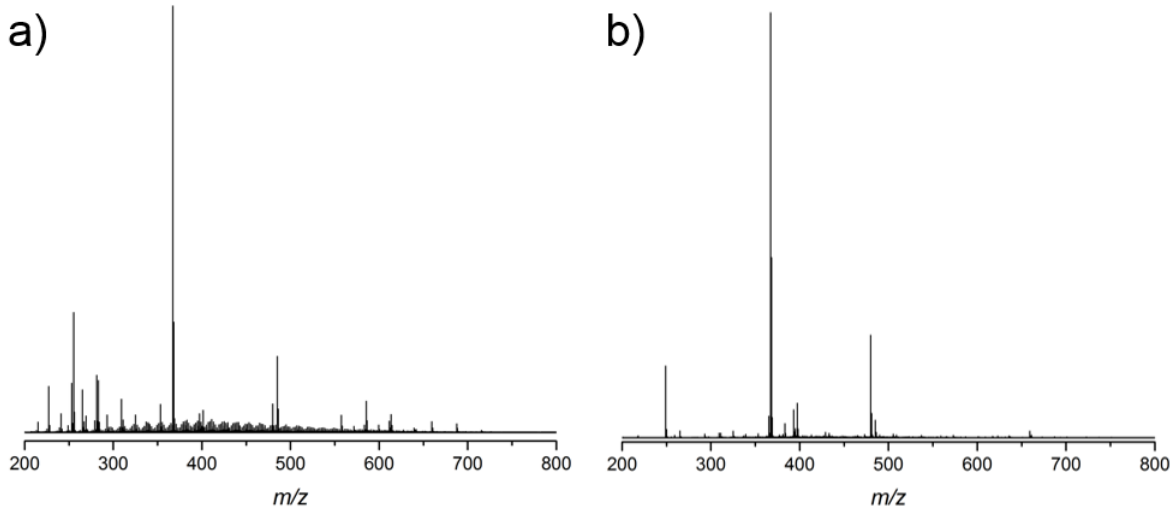


Figure 6.12 FTICR mass spectra of a) SRFA with CH_3 methylation and b) SRFA with CD_3 methylation. Large contaminant peaks are shown suppressing the sample signals.

The molecular masses of these large peaks identified them as belonging to silicon based compounds. The likely cause of this contamination was silicon grease from the vacuum/gas line that had pulled back into the system. Eliminating this presents a large

problem, as the silicone grease is used in a number of sections of our gas line setup to reinforce seals, including the Schlenk line, the vacuum controller and the rotary evaporator. In order to eliminate this contamination these systems might all require replacing, if they cannot be suitably cleaned and an alternative sealant to be used. Rather than attempt this problematic approach we decided to first test another possible way to remove the large peaks from the data. This was to use the internal quadrupole of the instrument to collect data "windows", these windows would be fitted around the large peaks, that would be excluded, and a better-quality data could be collected. A caveat associated with this is the loss of sample signals in the vicinity of these large peaks. However, the number of signals lost would is small in comparison to the number of signals currently being supressed, this can be seen in Figure 6.12 where only a few sample peaks are above the noise at each *m/z*. The calibration also becomes an issue as it means applying a calibration on windows to as opposed to an entire spectrum. It would still be possible to calibrate the spectra via external calibration however it will require identical windows of the SRFA standard to be collected alongside the SRFA-CD_3 windows. On a positive note a benefit is that by windowing the data we also increase the sensitivity of the peaks in the region analysed, resulting in increased signal to noise.

Five windows were collected at 8 MW, using Q1 masses and resolutions as described in Section 3.5.5.3, these resulted in windows from: 249 – 312, 318.8 – 364.8, 367.9 – 430.9, 431.5 – 478.5, and 480 – 600. The windows from the quadrupole are not hard limits and signals just above and below each range could be seen in each spectrum however their intensities were significantly reduced. This allowed the large

contaminant signals to be supressed whilst keeping the windows relatively close to one another so that fewer sample signals were lost.

Before going ahead with the full analysis of the data, we attempted to improve the data quality by acquiring at 4 MW. While there would be a reduction in spectral resolution, the higher consistency of data on a scan to scan basis should increase the signal to noise further. Collecting 4 MW increased the signal to noise of the sample but also of the contaminant peaks, because of this harsher window thresholds were needed in order to remove these peaks. Experimental details are described in Section 3.5.5.4 and produced windows covering: 255 – 345, 370 – 392, 400 – 426, 428 – 478, and 482 – 598, were collected. For comparison, the 428 – 478 window was also collected at 8 MW. Figure 6.13 shows an overlap of this window at 4 and 8 MW resolution. Although the signal to noise improvement was seen for 4 MW data the comparison of these windows shows that the resolution is not high enough as peaks that are resolved in the 8 MW spectrum have coalesced in the 4 MW spectrum.

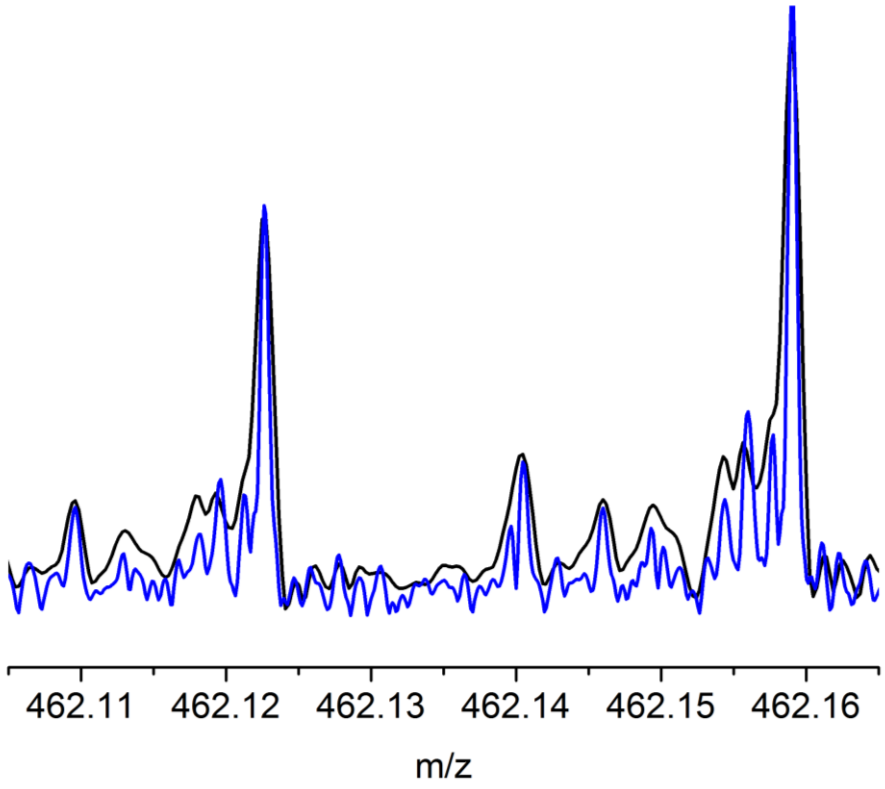


Figure 6.13 Overlay of 4 MW (black) and 8 MW (blue) spectra for the SRFA-CD_3 sample, showing that the lower resolution for 4 MW is insufficient to separate out the sample peaks.

6.9 Interpretation of FTICR mass spectra for methylated and deutero-methylated SRFA

This left us with the option of using the 8 MW previously acquired being the best in attempting to analyse the reaction on a real HS sample, these spectra are shown in Figure 6.14.

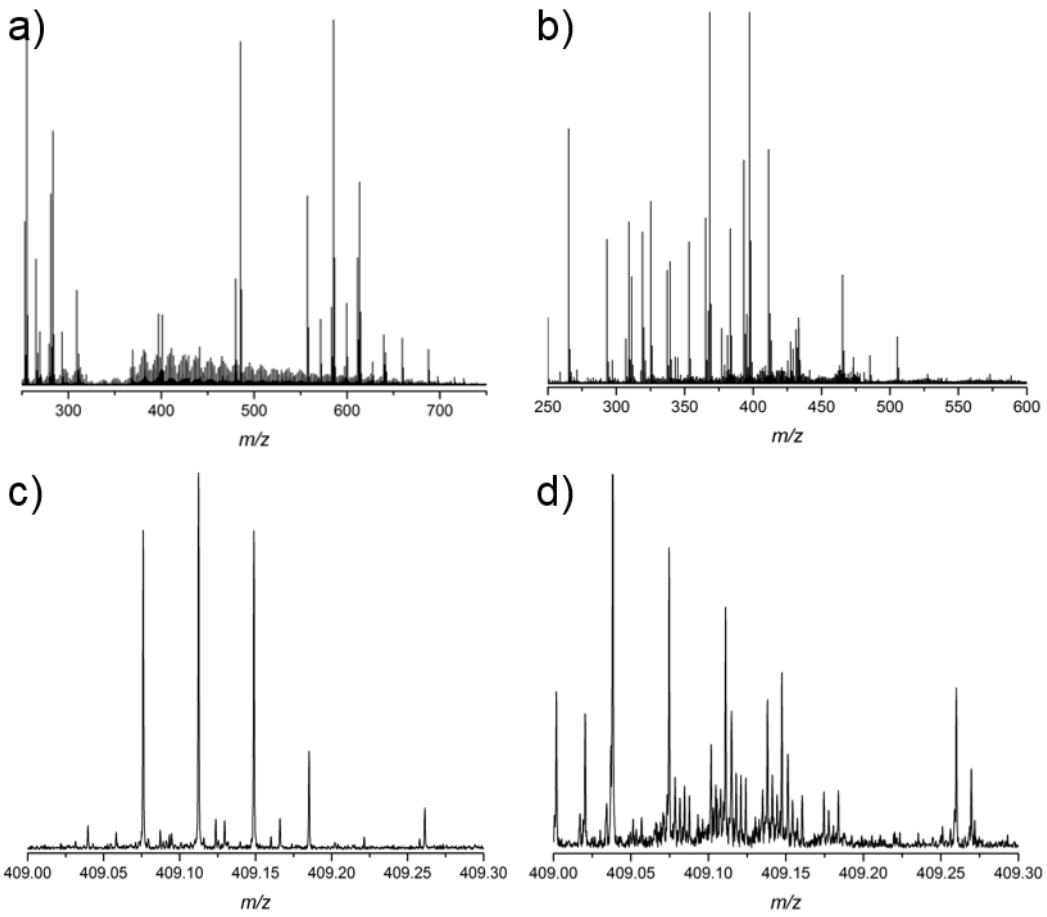


Figure 6.14 FTICR mass spectra of the combined windows for SRFA-CH_3 (a) and SRFA-CD_3 (b) alongside a single *m/z* for these two spectra (c and d, respectively).

Manual analysis of the sample peaks in Figure 6.14 reveals a trend, with peaks being regularly separated by ~1.5 mDa. Performing a purely formulaic comparison the difference, going from left to right, most closely fits a difference in formula between the two peaks of gaining two protons and losing a deuteron. As previously mentioned due to the inclusion of deuterium the assignment tools cannot accurately assign chemical formula as the difference in exact mass between potential formulae is smaller than the error threshold of assignment. Additionally, the prerequisite of assignment

being a thorough calibration, as discussing in section 3, is also an issue as the quality of calibration by applying a windowed calibration of another spectrum as an external is poorer by comparison to a standard calibration. To circumvent this and to reduce the impact of the calibration the SRFA-CH_3 data was assigned first. Cumulatively, 6605 peaks were identified, of these 2149 were assigned to monoisotopic formulae and 1038 were assigned to formulae containing at least one ^{13}C. Cumulatively this represents an assignment percentage of 48.3 %. While this is lower than we have seen for previous samples (see section 4), some of the unassigned peaks can be assumed to be contaminants brought in during the reaction, which can be seen in Figure 6.14. Using the 2149 monoisotopic formulae assigned as a starting point a secondary in-house code was developed in order to relate the SRFA-CH_3 formulae to the SRFA-CD_3 peaks. The software (Appendix 1) functions by taking each formula in turn and adding equivalents of 3.018831 Da, onto the mass of the formula, up to a maximum of 10, this takes the CH_3 methylated formula and predicts what its mass would become if it was deuteromethylated and the parent molecule contained between 1 and 10 carboxylic acid groups. These theoretical masses are searched for in the peaks identified from the SRFA-CD_3 spectrum, with an error allowance of ±0.375 mDa. If a peak is identified, it is recorded as being a methylated species, its number of deuterons is calculated and used to infer the number of carboxylic acid groups, and finally its parent formula is retroactively calculated. The result is a list of all identified formulae that contain carboxylic acid groups, including isomeric formulae that have varying numbers of groups. From the 2149 species identified in the SRFA-CH_3 spectrum, comparing theoretical peaks to the 9965 peaks identified in the SRFA-CD_3 spectrum, a total of 1723 COOH containing compounds were identified. These species ranged from 1 to 5

carboxylic groups and the number of formulae identified of each can be seen in Table 6.2.

Table 6.2 Number of formulae identified that contain carboxylic acid groups alongside the type of molecules as determined by $AI_{(mod)}$

# of COOH	# of aliphatic formulae	Number of aromatic formulae	# of condensed aromatic formulae	# number of formulae
1	515	129	14	658
2	388	199	47	634
3	118	158	47	323
4	4	69	25	98
5	1	4	4	9

Also included in the table is the number of the each type of formulae identified, aliphatic, aromatic and condensed aromatic, for the parent compound as calculated by $AI_{(mod)}$, described in Section 5. From this we see a distinct change in molecular type as the number of carboxylic acid groups increased. Aliphatic molecules, that made up over three quarters of the formulae contain one carboxylic acid are potentially missing for 4 or more carboxylic groups. The trend is no so steep for aromatic and condensed aromatic, which although also reduced in number at higher carboxylic numbers, make up a large percentage of all peaks identified. It is difficult to decide if this is a true reflection of the abundance of these compounds in the sample or a consequence of different ionisation potentials. By removing the labile proton of the carboxylic acid we make the molecule much less ionisable, which means by methylating 4 and 5 carboxylic acid groups we are drastically reducing the ionisation potential of the molecules. Thus, to still be ionised by ESI, the molecule must contain a number of

additional oxygens in alcoholic or phenolic environments, this means the remainder of the molecule will have to be more condensed to allow for this. So aliphatic molecules may, and probably are, present with 4 or more carboxylic groups however they will likely lack the ionisation potential to be seen by this method.

Van Krevelen diagrams for the assigned formulae are given in Figure 6.15, points associated with 5 carboxylic acid groups have been omitted, due to there being too few point to reasonably produce the kernel density estimation (KDE) plots or compare trends from the points on the van Krevelen diagram.

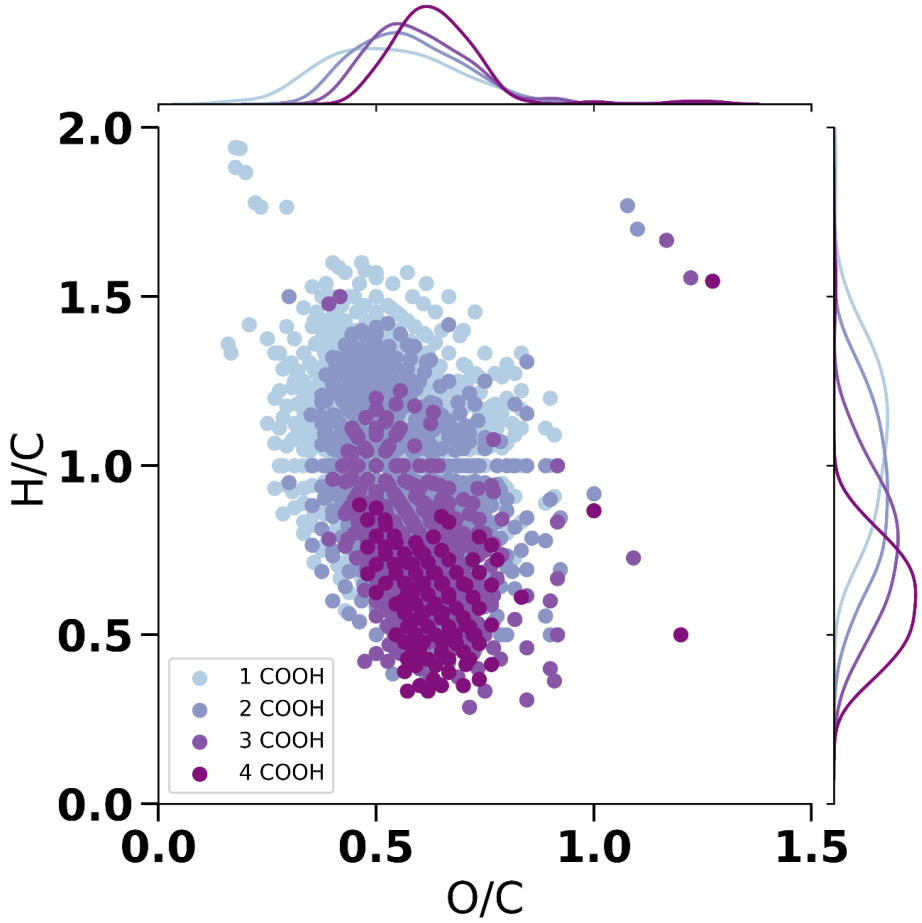


Figure 6.15 Overlaid van Krevelen diagrams of formulae identified with 1 – 4 carboxylic acid groups. KDE plots for these data sets attached for both axis.

Comparing these assignments in this manner allows for trends to be seen in both the average H/C ratio and O/C ratio. As number of carboxylic acid groups increase the H/C ratio decreases, and the O/C ratio increases. Looking firstly at the O/C ratio, this increase fits with what we would predict, because as the number of carboxylic acid groups increases the number of oxygens in methyl esters increases, from 2 oxygens for 1 carboxylic acid group to 8 oxygens for 4 carboxylic acid groups. These methyl ester oxygens aren't acidic, therefore additional alcoholic oxygens are required for ionisation by ESI-(-). Overall this means the more carboxylic acid groups a molecule contains, and on average the more oxygens a molecule will contain overall in order to still be ionised by ESI, thus the trend we see in Figure 6.15. The H/C decreasing trend can be associated with the instrumental parameters combined with the chemistry of the molecules identified. Treating the problem in a mathematical approach, the mass range we identify these compounds over has the same limits for all species (300 – 600 *m/z*), meaning when we compare species like this we are comparing species of a relatively similar mass to one another. Mathematically, a carboxylic acid group ($-CO_2H$) by itself has a nominal mass of 45 g/mol^{-1} and a H/C ratio of 1, whereas different groups with similar masses, for example, have; propyl ($-CH_2CH_2CH_3$), 43 g/mol^{-1}, H/C ratio of 2.33; ethanal ($-CH_2COH$), 43 g/mol^{-1}, H/C ratio of 1.5; and ethanol ($-CH_2CH_2OH$), 45 g/mol^{-1}, H/C ratio of 2.5. Comparing these groups we can see that the H/C ratio of a carboxylic acid group is relatively low compared to its mass. This can be used to explain what causes the H/C trend in Figure 6.15, as the number of carboxylic acid groups increases, whilst keeping a formulae with a constant average mass, mathematically and therefore formulaically we lose atoms that would originally increase the H/C ratio of the molecule, this reduces the molecules overall H/C ratio.

6.10 Comparison of results and experimental caveats

Although we can successfully identify CD_3 containing compounds and find parent formulae for over 1700 carboxylic containing species there are still a large number of peaks in the spectrum that aren't accounted for by this analysis. Some will belong to CHO series of molecules that contain no carboxylic acid groups and thus weren't methylated however there are more unidentified peaks than CHO species could account for. Analysis using the aforementioned software was repeated with an alteration to identify CHD_2 series, following which, manual analysis of the remaining peaks, using software that generates all chemically possible molecular formula within a predefined range, in a 0.1 *m/z* range (Figure 6.16) revealed they belonged to D_1, D_3 and D_5 species. Of the 23 peaks in this 0.1 *m/z* range four were still unassigned, these peaks had too many possible formulae so a confident assignment could not be achieved. From the manually assigned formulae: the D_1 species are likely due to CH_2D methylation; the D_3 species were potentially unassigned due to either their CH_3 methylated counterpart not being assigned in the SRFA-CH_3 spectrum, or the peak being outside assignment thresholds for the assignment program; the D_5 species could be formed by a combination of CD_3 and CHD_2 methylation of a molecular species with two carboxylic acid groups.

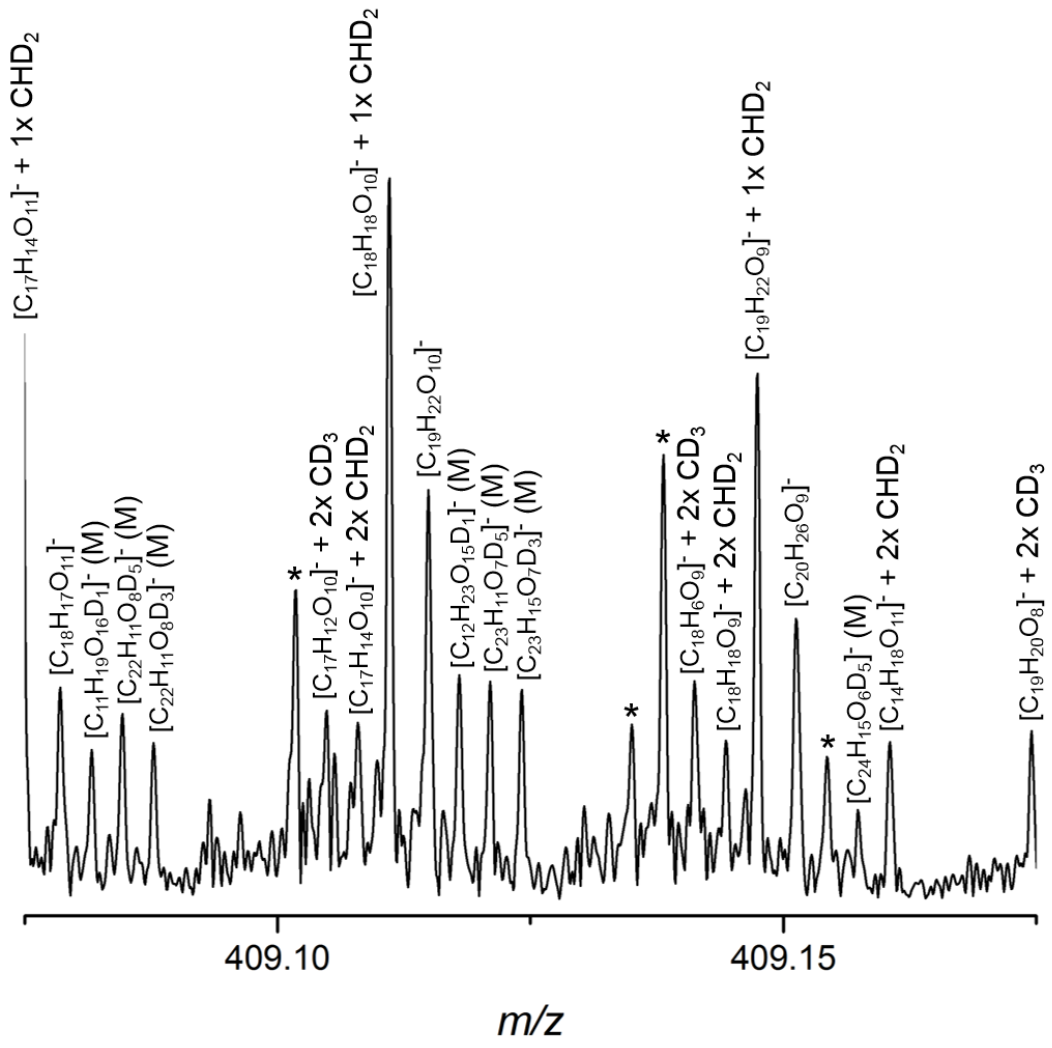


Figure 6.16 SRFA-CD_3 window (Figure 6.14) focussed into most complex 0.1 *m/z* region. Assignments shown based on unmodified formulae as well as parent formulae of CHD_2 and CD_3 methylated series. Manual assigned peaks denoted by (M) and unassigned peaks by (*).

This indicates that despite the model compounds showing much lower amounts of CHD_2 by-products, in the HS sample, despite the sample pre-treatment, these were being formed in a similar, if not greater, abundance to the CD_3 compounds. A likely

cause of this being worse for the real sample than for the model compounds is the proton to deuteron exchange on the carboxylic acid site of the SRFA itself, (see Figure 6.4). Subsequent titration experiments showed that the labile protons in HS samples were held more tightly, possibly in aggregates or clusters, and thus were released at a much slower rate. The titrations showed this as the pH of titrated could take from several hours up to a day to stabilise upon addition of base. This would account for the severe drop in CD_3 product over CHD_2 and CH_2D. Ultimately this leads to the previously discussed problem concerning false positives: is a formula x with D_6 genuinely 2x D_3 reporting a 2x COOH or is it 3x D_2 reporting 3x COOH and incorrectly assigning the number of carboxylic acid groups.

An additional problem with the experiment is the signal to noise of the peaks in the sample, as can be seen from the snapshot in Figure 6.16. This signal to noise is very low and means we potentially lose a large number of peaks into the noise. There are a number of causes for this; the production of D_1 and D_2 by-products, but also the nature of the reaction itself. Three factors of the reaction decrease the signal intensity in the spectrum, the contaminants, the removal of labile protons, and the separation of isotopes. The first two of these have been discussion previously but to recap, the large peaks caused by contaminants supress the sample signals and the removal of carboxylic protons heavily reduces the ionisation efficiency of the molecule. The third effect is due to the complexity of HS, each peak in the spectrum is a single formula, but is almost certainly originates in several isomers; this is a known issue for mass spectrometry as it is unable to separate out these isomeric peaks using standard methods. When we methylate the sample, we cause these isomers to be separated out, based on the number of carboxylic acids each contains. The groups of isomers for each

number of carboxyl groups are moved to a new mass on the spectrum, where no other peaks fall, and thus each peak is representing a smaller number of compounds and will have a lower signal intensity. Combing these issues, it makes the data produced by this experiment questionable, and related heavily to a previous statement that a reaction on an unknown complex mixture must be fully understood and work perfectly in order to produce usable quality data. The experimental approach we have tested seems to be unsuitable for the task, as its desired yield of CD_3 containing product on our complex system was not high enough to produce confident results.

Similar work has recently been published by Zherebker *et al.*[106] where carboxylic functionality was counted aided by an alternative methylating agent, thionyl chloride. The reaction itself involves a 4 h reflux, which is a much harsher treatment than we used. We selected the TMS-diazomethane reaction due to it being such a gentle reaction; the danger with doing any chemistry on an unknown mixture is that it is impossible to tell if the mixture was altered in an unexpected way, *e.g.* hydrolysing esters or oxidisation. Reflux in an open-air environment is therefore not suitable for HS modification. Their method also included an extraction step post reaction using a solid phase extraction cartridge; this allowed the authors to remove any contaminants left over from the reaction. This was a step our protocol did not include, and could have the potential to remove some of the contaminants from our sample, providing they do not co-elute. Another difference between the two approaches was that unlike our work the authors did not produce a CH_3 and a CD_3 sample for assignment, instead they prepared only a CD_3 sample. As a result, they were unable to assign the spectra using the software available to them and had to resort to manual peak assignment. Using this method, they assigned the 500 highest abundance peaks and identified

species containing from 0 to 6 carboxylic acids. While this generally agrees with what we observed, their manual assignment was questionable, as they used an error threshold of 0.001 *m/z* across the entire mass range, this is three times larger than the threshold we used and represents an error margin of 2.5 ppm at 400 *m/z*. This error margin is very high, considering the spectra and data produced from this modification is several times more complex than the original sample, therefore we argue that such large error thresholds would be too unreliable. Their reaction also appears to have produced CHD_2, or similar species, and hasn't achieved quantitative yields, however as they have only focused on the 500 most abundant peaks they have been able to ignore these lower lying by-products. Of the 500 peaks assigned, 256 were assigned to species containing at least one carboxylic acid group. Table 6.3 shows a comparison of number of formulae assigned for each number of carboxylic acid groups present.

Table 6.3 Number of formulae assigned as containing from 1 – 6 carboxylic acid groups for SRFA by the two different approaches.

Number of COOH	Our approach	Zherebker's approach
1	658	48
2	634	37
3	323	87
4	98	73
5	9	10
6	0	1

Comparing both approaches it is clear that molecules with 5+ carboxylic acid groups are much less abundant as per these methods. This could be due to a difficulty in ionising these compounds once they have lost their most acidic protons, however, it also could provide information about the structure of molecules present and indicate that species with 5+ carboxylic acid group are significantly less common. Both methods fail to identify enough compounds to confidently confirm the presence of molecules containing 5+ carboxylic acid groups, 9 from 9965 peaks for our method, and 11 from 11094 peaks for Zherebker's method. The maximum number of formulae identified was for species containing 1 carboxylic acid group for our approach and 3 carboxylic acid groups for Zherebker's. While it is impossible to conclude which result is correct with completely certainty, it would follow that our analysis here is superior due to it taking into account all peaks in the spectrum, not just the 500 most abundant. This truncation of their data prevents any reliable trends from being drawn, which can be seen from their number of formulae assigned dropping between 3 and 2 carboxylic acid groups but then increasing between 2 and 1 groups. It also follows that species containing one carboxylic acid group would be more easily identified due to it involving the smallest number of acidic protons being removed from the molecule, thus having the lowest effect on ionisability.

Considering both approaches, we can see flaws and faults with both reactions. Both suffer from incomplete deuteration, despite much harsher reaction conditions used by Zherebker, which coupled with the increase in complexity of including deuterium in formula calculation software, makes assignment significantly harder. Our approach to assignment using a CH_3 sample to reduce the mass difference from 17 Da down to 3 Da per COOH allowed us to more reliably assign formulae as it reduced the effect of

the calibration drifting the further we move up the spectrum from the starting formula. This approach to analysis worked well and was an effective method for assignment.

A major problem with the analysis itself was using ESI as the ionisation source, as the nature of the reaction makes the product less ionisable by ESI. Therefore, the more a molecule is methylated the less ionisable it becomes, meaning the more interesting species containing large amount of carboxylic acid groups are at the most risk of not being detected by this method. With that in mind it would be might be more appropriate to use other ionisation sources like LDI and APPI in order to perform this analysis. In order to successfully test that however a more reliably level of deuteration would be required, whether by improving the efficiency of the reactions presented here or by using a different reaction, in order to confidently assign formulae and carboxylic acid numbers in this manner the reaction needs to work almost perfectly with negligible levels of interference. Chemical reactions are rarely so controllable and predictable, however due to the immense complexity of these systems such modification reactions are not just idealised but required.

Chapter 7: Isotopic labelling of labile hydrogen sites via carbon-13 enriched methylation for NMR studies of HS

Isotopic labelling also has applications as a "spectroscopic" separation technique for NMR studies of HS. A heavy isotope of carbon (^{13}C) is NMR active whereas its natural form ^{12}C isn't. Natural abundance of ^{13}C is ~1 %, meaning a site that has been selectively labelled with this heavy isotope will have a 100 fold sensitivity increase with regard to its natural counterpart. Sensitivity is one of the major factors for NMR when it comes to analysis of complex mixtures with many low abundance species, and by improving sensitivity we could open the door for a wide range of advanced NMR experiments. Introducing a 100% labelled ^{13}C allows us to select only the molecules that carry this isotopically enriched group. With this goal in mind a standard methylation reaction was tested using methyl iodide. Initial exploration of this reaction was performed by Bell *et al*.[88] on a model mixture of phenolic compounds, however due to the significant difference in behaviour of an actual HS sample the method required investigation for the use on humic substances. The reaction scheme for this reaction is shown in Figure 7.1. The main aim of these modifications was to create a final sample that was reliably labelled with $^{13}CH_3$ groups at labile hydrogen sites (mainly alcohol and carboxyl sites).

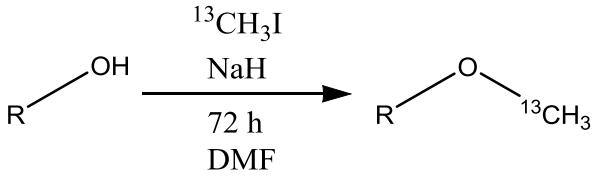


Figure 7.1 Reaction scheme for the methylation of labile proton sites using MeI.

7.1 Methylation of Red Moss humic acid

A sample of Red Moss humic acid (RMHA) was extracted in accordance with IHSS guidelines by A. Michalchuk. This sample was selected due to its locational significance and as it contained a greater amount of aromatic molecules. The initial reaction was performed using 50.3 mg of RMHA, 42 equivalents of NaH and MeI, ran for 72 h, quenched using 0.1 M HCl and 20 cm^3 water, following which it was worked up using $CHCl_3$ washes and dried over $MgSO_4$. During the reaction a dark particulate matter could be seen within the reaction vessel, and upon work up ~5 mg of product was collected, suggesting that the HA hadn't successfully dissolved in the solvent system or had crashed out during the reaction and was lost when the $MgSO_4$ was filtered off.

7.2 Methylation of a model compound

Of the two aforementioned possibilities the latter was tested first, in order to circumvent this step the work up would need to be made water free, to remove the necessity for filtration that would remove undissolved matter. To do this the quenching method was altered to using anhydrous MeOH instead of HCl/H_2O and the work up was changed to an extended evaporation period to 8 hours. For clarity, the original workup will be described as the "classic workup", and this new method as the

"evaporation workup". A single model compound (**8**, 50.6 mg) was used to test the reliability of the method alterations. The major concern with the evaporation workup was that it wouldn't remove the salt by-products; NaI and NaOH; which could adversely affect the quality of the NMR spectra. Upon evaporation a much larger amount of material was produced however due to the salt the exact yield could not be established. Figure 7.2 shows a ^{1}H NMR spectrum of the product, the fully methylated product can be seen as the major product with very little starting material remaining, this confirms the activity of the reagents. However the spectra is dominated by DMF solvent peaks, a side effect of the increase in material formed is that the residual DMF was significantly more difficult to remove, even after increasing the evaporation time 4-fold. This was seen as a minor issue due to the possibility of DMF removal to be improved via further evaporation or by exploitating DMF forming an azeotrope pair with toluene.

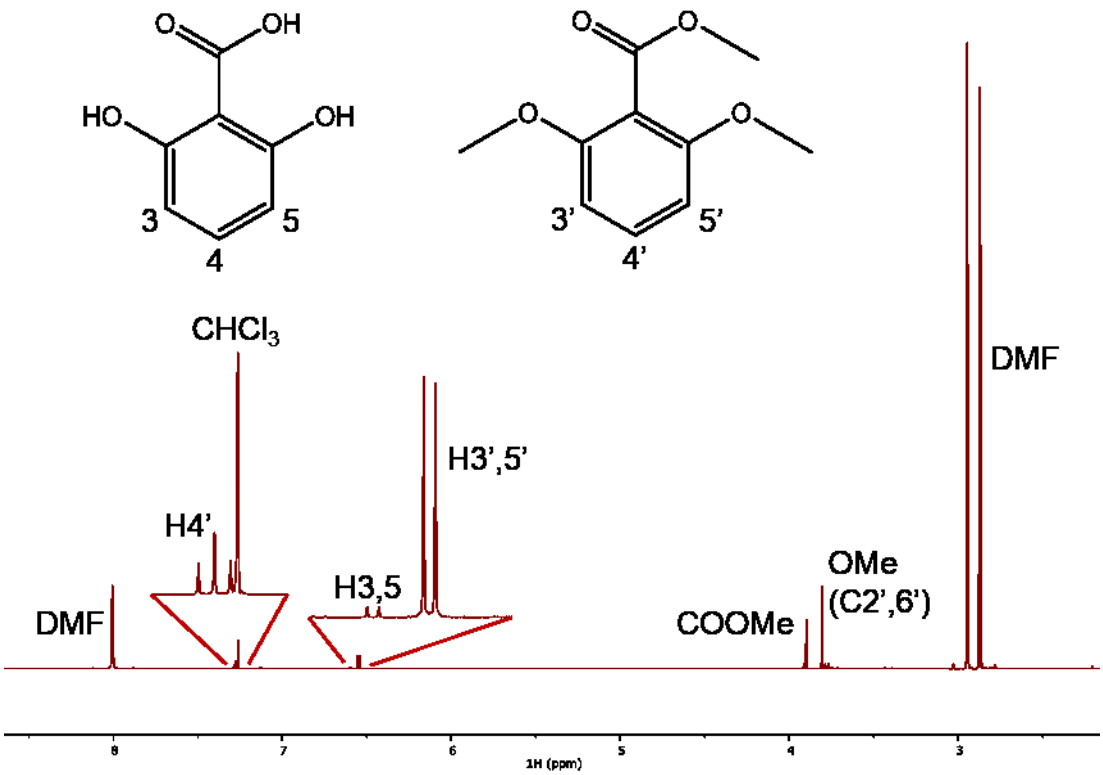


Figure 7.2 ^{1}H spectrum of the reaction mixture of methylation of 2,6-dihydroxybenzoic acid (**8**). The inset shows aromatic signals of the product (major peaks) and starting material (minor peaks). H4 proton not seen due overlap with the signals of H4' and $CHCl_3$ protons.

7.3 Methylation of RMHA using a new method

56.2 mg of RMHA was used to test the reaction again with the evaporation workup, additional evaporation steps with 10 cm^3 toluene being added, the flask shaken and then removed were performed after the final evaporation from the previous reaction. Three consecutive washes were performed and NMR spectra gather before and after showed a DMF:$CDCl_3$ concentration ratio drop from 4:1 before to 1.3:1 afterwards. Three more washes were done before the product was collected. Unlike the model compound the methylation of RMHA produced two products, crystals that had

collected at the bottom of the round bottomed flask and a fine powder that coated the crystals and the walls of the flask. Both products were an off-white/very light brown colour. The total amount of material became ~400 mg, most of which will be salt produced by the reaction. It was theorised that these crystals were salt crystals formed in regions of high salt concentration and the powder was a mixture of salt that hadn't formed large crystals as well as the methylated RMHA product. In order to confirm this theory the two product forms were separated, 100 mg of each was taken, dissolved in $CDCl_3$, centrifuged and filtered through a plugged pipette to product an NMR sample. The 500 MHz NMR spectra were compared (data not shown), and whilst both contained RMHA peaks, their intensity was significantly higher in the powder sample than the crystal, suggesting our theory was correct. The residue remaining in the pipette from the powder sample was then washed with DMSO-d_6, agitated, separated and filtered. The NMR spectrum of this sample was collected and compared to the one obtained from the $CDCl_3$ sample. The DMSO-d_6 spectrum showed the same product peaks as the $CDCl_3$ spectrum without any new peaks being identified; suggesting $CDCl_3$ was successfully dissolving the product formed. The remaining powder product ~200 mg was washed with the $CDCl_3$ sample as described previously in an attempt to increase the concentration of the sample. This ^1H spectrum was collected at 800 MHz, where an initial spectra were collected, to estimate its concentration and evaluate the outcome of methylation (Figure 7.3).

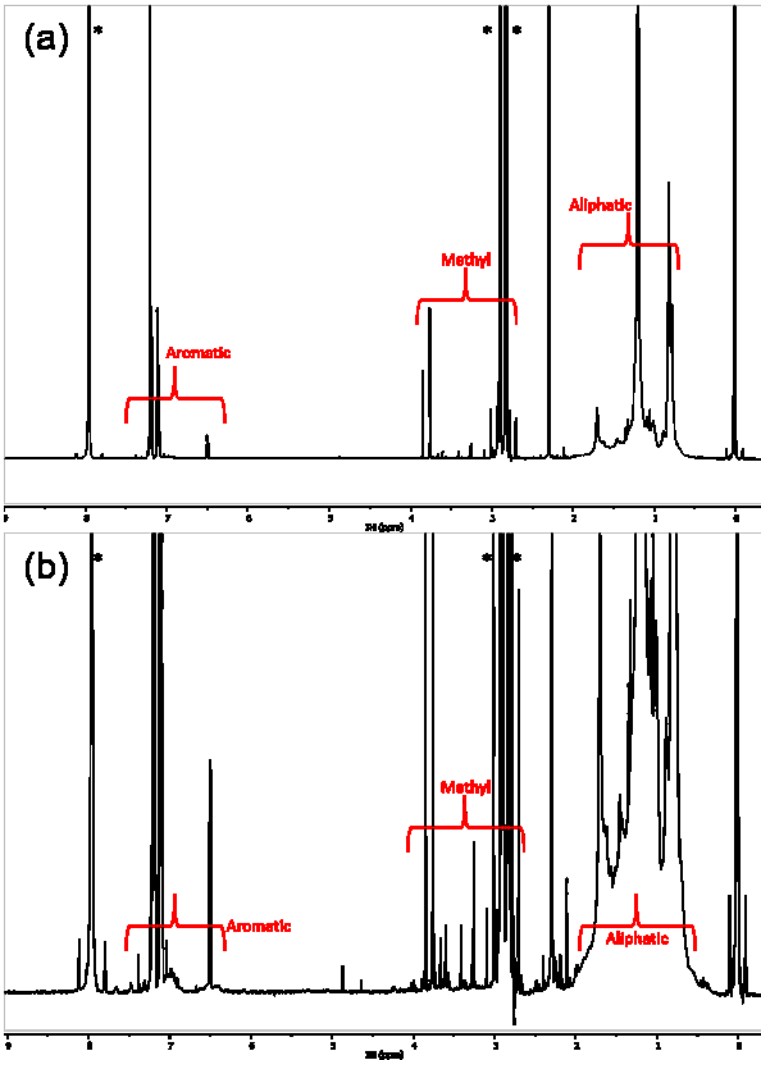


Figure 7.3 ^{1}H NMR spectrum of methylated RMHA, (b) represents a vertical expansion of (a). Residual DMF is denoted by *; (b) shows the RMHA peaks at low intensity.

The spectrum showed that the extraction was successful as peaks associated with the sample are present, *e.g.* aliphatic resonances at 0.5 – 2 ppm. The region around 4.0 ppm contains methyl singlets indication some methylation, however these are still very

low intensity and only few. Overall, the spectrum is lacking the true complexity associated with HS. This indicates very low efficiency of HA methylation. Considering the high yield of the reaction on the model compound as shown in Section 7.2, the issue must clearly be with the sample being used. HA is notoriously difficult to solubilise in anything other than alkaline water, and due to water reacting violently with NaH, water is not a suitable solvent for this reaction. In an attempt to further investigate this issue, we decided to perform the reaction on another type of HS, fulvic acid. The fulvic acid fraction of HS is usually much more soluble, hence this sample is indeed suitable to test if sample solubility has been the limiting factor.

7.4 Methylation of SRFA

Upon addition of SRFA into the reaction vessel, the improvement in solubility was instantly apparent with the solution producing a deeper and more opaque brown colour, than had been seen with the HS sample. As the reaction progressed the familiar white crystals formed, as discussed in Section 7.3. A noticeable change occurred when methanol was added in order to quench the reaction, a colour change occurred as the solution went from a light brown to a darker brown. Everything else progressed as expected, including the formation of large quantities of salt crystals, and with DMF removal continuing to be a difficult and laborious task.

The powder residue was dissolved into three $CDCl_3$ samples which were then concentrated into a single sample to be analysed on the 800 MHz spectrometer. Initial spectra indicated high salt concentration, as the sample could not be correctly shimmed, which resulted in poor quality data. White powder was noticed crashing out of the sample. The sample was passed through a 0.45 μm filter, left overnight in a

fridge at 5 °C to allow more salt to crash out and filtered again. Following this the sample could be shimmed. A ^{1}H, ^{13}C-HSQC spectrum (data not shown) indicated the presence of methyl oxygen peaks in the ester, aromatic and alcohol regions. This result indicated that the solubility of the sample is important. The 2D spectra contained intense t_1-ridges of DMF, which was still at too high a concentration. The sample was therefore treated to further evaporation steps to remove more DMF. During evaporation salt continually crashed out from the solution suggesting that residual salt was dissolved in the DMF. The treated sample showed an approximately 64 fold reduction in DMF concentration and the desired signals could be seen with much more clarity, Figure 7.4, with the methyl signals now prominent within the spectrum.

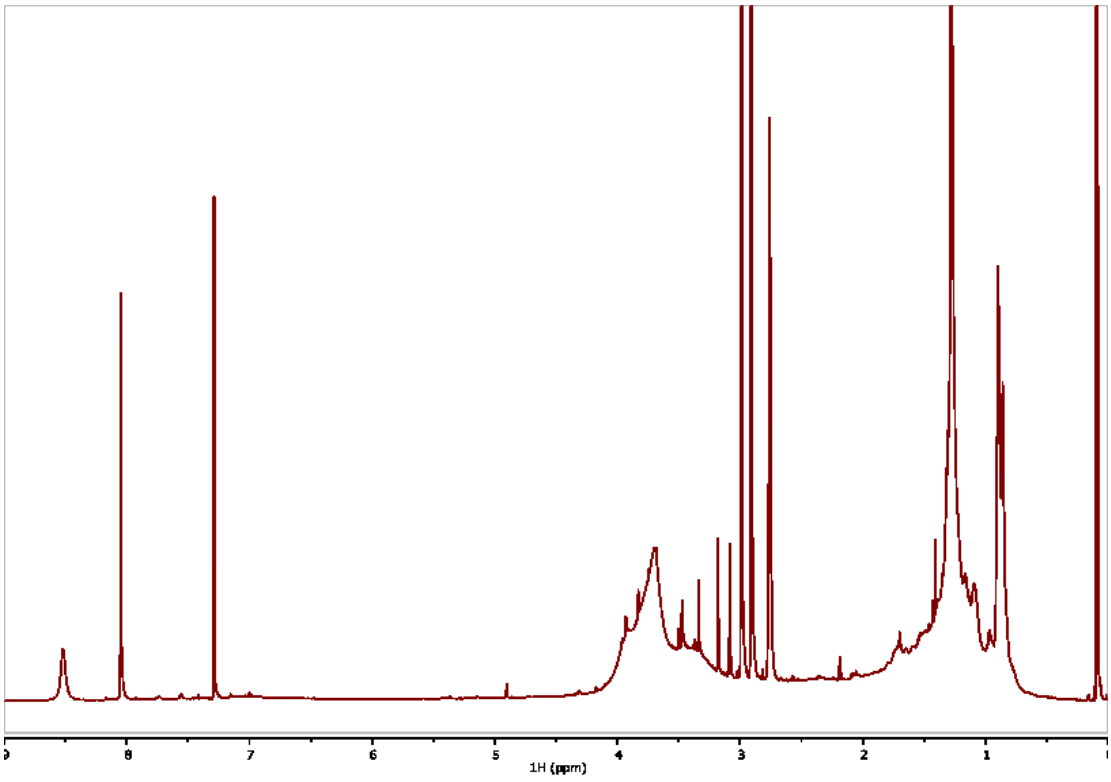


Figure 7.4 ^{1}H spectrum of methylated SRFA. Methyl signals due to methylation now have a significant intensity in the 3 – 4 ppm region of the spectrum (compared with Figure 7.2). This confirmed successful methylation of the sample.

Ultimately, the lack of methylation of HA is caused by the poor solubility of HA in a purely organic solvent system, an issue that is not easily resolved, and its counterpart FA was able to solubilise in the solvent system, thus was successfully methylated. The major challenge left to face was efficient DMF removal, which took us back to considering the workup stage of the reaction. Once solubility was identified as the issue, the classic workup was considered again, as the $MgSO_4$ filtration section of the workup should have no issues with a soluble sample.

7.5 Methylation of Red Moss fulvic acid

Next, we investigated a Red Moss fulvic acid (RMFA) sample, as this sample was from the location being investigated in our group. The reaction was repeated in the standard fashion using 33 mg of RMFA until quenching, at which point the reaction was separated into two equal volume fractions; the first was quenched using methanol, similarly to the experimental of the evaporation workup; the second was quenched by drop wise addition of concentrated HCl, according to the classic workup. The light to dark brown colour change as seen for SRFA was seen during both the quenching setups and used as an indication of the quench being complete. The first fraction, the evaporation workup, was subjected to evaporation for 15 h at 60 °C and 18 mbar, and the second, the classic workup, was processed using $CDCl_3/H_2O$ washes, followed by drying over $MgSO_4$ and evaporation for 1h at 60ºC and 18 mbar. Both the evaporation and classic workups removed a similar amount of DMF, the classic workup proved to be a much faster a reliable method. It also benefits hugely from the removal of salt from the final product due to it being extracted into the aqueous phase of the workup.

While the exact yield cannot be measured due to the remaining DMF, which was not successfully removed, approximately 6 mg of product was collected. Saturated NMR samples were made from both producta and compared to one another. While both methods successfully removed a similar amount of DMF, for the pure evaporation method, the intensity of the product peaks were ~50% and ~35% lower for the methyl and aromatic regions respectively, compared to the complete work up. We are unsure whether this is due to product loss during the extensive evaporation procedure or the salt presence lowering the saturation threshold. Regardless it points to the complete work up being much more suitable due to the product concentration increase and due to the easy removal of salt preventing it from stopping the spectrometer from correctly shimming.

7.6 ^{13}C enriched methylation of RMFA

With the reaction and the workup for FA, it was ready to be taken on and used for isotope labelling, as was the initial aim; this method is described in Section 3.3. To achieve sufficient signal to noise in higher dimensional spectra a more concentrated sample would be required, approximately 10 – 20 mg in a 600 µL sample. Due to the lower yield from the previous reaction, 6 mg from 16 mg, it was decided to use a higher starting mass of ~60 mg. ^{13}C-methyl iodide was used as the source of ^{13}C which results in isotopically labelled methyl esters and methyl ethers being formed. The reaction progressed as expected and nothing different was observed in comparison to the SRFA methylation. A ^{1}H, ^{13}C HSQC spectrum was acquire on the 800 MHz spectrometer, (Figure 7.5). Here was see a large sensitivity increase associated with using the ^{13}C label producing clusters of cross peaks in the methyl region. At the same time, there

are some significantly stronger peaks that stand out above the complex "humps". These peaks must belong to chemical species that are either dominant within the RMFA or that are more efficiency methylated, including the efficiency of their interaction during workup.

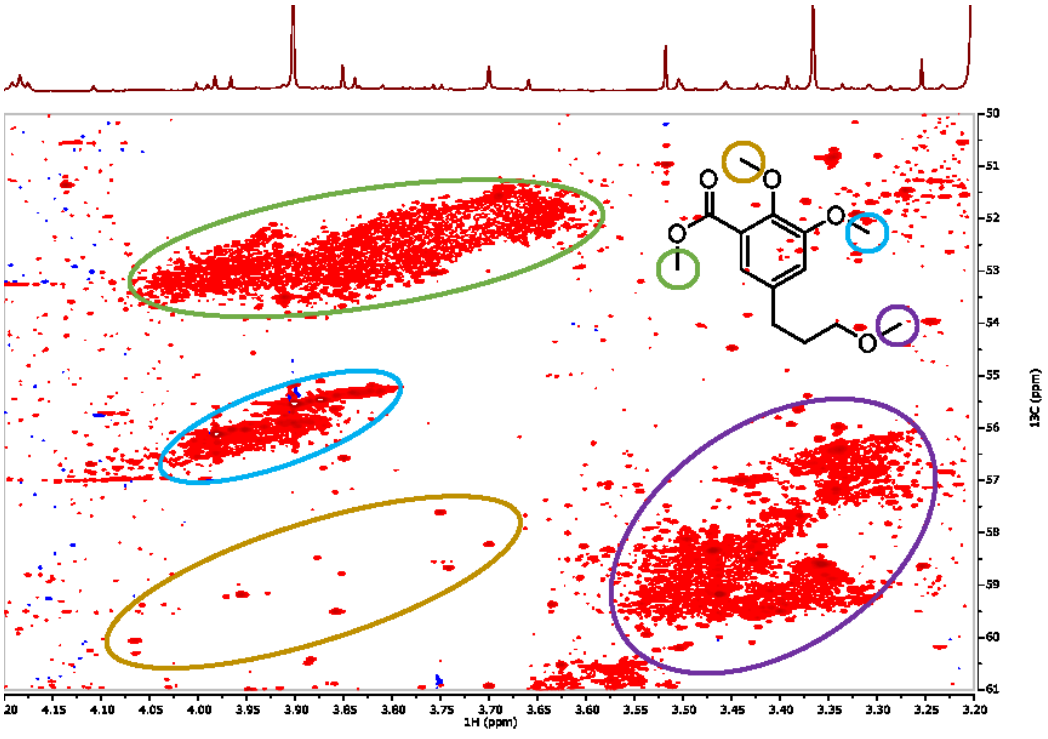


Figure 7.5 A 800 MHz ^{1}H, ^{13}C-HSQC spectrum of ^{13}C methylated RMFA with four distinct regions highlighted alongside a hypothetical structure showing MeO types. The regions are methyl esters (green), phenolic methyl ethers (blue), aliphatic methyl ethers (purple) and "sandwiched" aromatic ethers (gold).

The spectrum contains clusters of signals, belonging to carboxylic methyl ester, phenolic ethers, aliphatic ethers and "sandwiched" ethers, all of which are highlighted. The sample was passed onto Dr N.G.A. Bell to be analysed by higher dimension 3D and 4D NMR experiments.[37] *Bell et al.* used these experiments to produce the first structures of aromatic compounds within HS obtained by NMR.

Chapter 8: Conclusions

Humic substances (HS) are naturally formed complex mixtures with a wide range of geochemical functions, which makes them a highly interesting subject for environmental studies. HS potentially contain tens of thousands of individual species that cannot be fully separated by traditional methods. This makes HS a massive challenge for modern analytical chemistry. High-resolution nuclear magnetic resonance (NMR) spectroscopy and Fourier transform ion-cyclotron resonance mass spectrometry (FTICR MS) hold the best promise for solving the chemical structures of compounds imbedded in these mixtures. However, both techniques are currently unable to provide full structural characterisation. We have investigated the potential of these methods in improving the information gained on HS samples.

Tools for the automated analysis of FTICR MS data were evaluated as this is one of the initial steps of any investigation. We found that a number of factors were important in the choice of the software used. The major factors were the running time, the accuracy of the assignment and the maximum size of data sets that can be analysed using the software. Comparing two types of software, KMDv_0.5 and PetroOrg, we observed that PetroOrg was more suitable; it returned more accurate assignments on larger data sets.

The choice of the ionisation technique of HS for FTICR MS studies was scrutinised next. Laser desorption/ionisation (LDI), a mostly abandoned ionisation method, has proven not to cause fragmentation, based on comparison with a trusted related method, matrix-assisted laser desorption/ionisation (MALDI). LDI and MALDI were compared to electrospray ionisation (ESI) on the analysis of Suwannee River fulvic

acid (all three) and Suwannee River natural organic matter (LDI and ESI only). Using heteroatomic class distributions and aromaticity index, it was seen that ESI primarily ionised aliphatic and higher oxygen number compounds, whereas LDI/MALDI ionised more aromatic and lower oxygen number compounds. From this work we determined that both ESI and LDI should be used in unison for analysis of HS in order to provide a more complete representation of the sample.

Focussing on ESI, a widely used methodology for the analysis of HS by FTICR MS, various aspects of the methodology were tested. Concerns, such as collision cell voltage dependent fragmentation and aggregation at higher sample concentrations were shown not to occur within standard operating procedures on our Bruker 12 T SolariX FTICR mass spectrometer. Notably, we found that the 200 – 700 *m/z* mass range currently associated with ESI FTICR MS of HS, is an incorrect representation of these mixtures. By tuning the spectrometer to select the low or high mass species, separately, we were able to detect molecular ions ranging from 109 *m/z* to 1035 *m/z*. A comparison of the species present in this wider *m/z* range using aromaticity index, showed increased chemical diversity at lower masses (<300 *m/z*).

Isotopic labelling was investigated as a basis for analysis of key functionalities within HS using MS and NMR. TMS-diazomethane was used in conjunction with FTICR MS. TMS-diazomethane allowed for selective deutero-methylation of carboxylic acid groups. The isotopic labelling was shown to be successful for model compounds, however, on HS samples the reactions performance deteriorated, producing a mixture that was significantly more difficult to analyse. Regardless, automated software for the analysis of complex deutero-methylated spectra was developed and used to aid

enumeration of the chemically significant carboxylic acid groups. We were able to identify parent formulae that contained between one and five carboxylic acid groups.

Methyl iodide was used to incorporate ^{13}C methyl groups into HS molecules for NMR spectroscopy studies. Model mixtures and HS samples were methylated in order to prepare samples required for the development and application of complex multi-dimensional NMR experiments for structural characterisation of aromatic HS molecules.

In summary, studies of HS in terms of the structure of their constituent molecules are challenging. The work presented here shows positive developments in the application of NMR spectroscopy and FTICR MS combined with isotopic labelling for this field. We have tested tools and methods for the analysis of HS by FTICR MS and NMR, as well as expanded, and in some case corrected, the technical knowledge required for proper use of FTICR MS for the analysis of HS.

Bibliography

1. Batjes, N. H., Total carbon and nitrogen in the soils of the world. *European Journal of Soil Science* **2014,** *65* (1), 10-21.

2. Schmidt, M. W. I.; Torn, M. S.; Abiven, S.; Dittmar, T.; Guggenberger, G.; Janssens, I. A.; Kleber, M.; Kogel-Knabner, I.; Lehmann, J.; Manning, D. A. C.; Nannipieri, P.; Rasse, D. P.; Weiner, S.; Trumbore, S. E., Persistence of soil organic matter as an ecosystem property. *Nature* **2011,** *478* (7367), 49-56.

3. Fontaine, S.; Barot, S.; Barre, P.; Bdioui, N.; Mary, B.; Rumpel, C., Stability of organic carbon in deep soil layers controlled by fresh carbon supply. *Nature* **2007,** *450* (7167), 277-280.

4. Leenheer, J. A.; Croué, J.-P., Peer reviewed: characterizing aquatic dissolved organic matter. ACS Publications: 2003.

5. Tan, K. H., *Environmental Soil Science, Third Edition*. CRC Press: 2009.

6. Lowe, L. E., Chapter 2 Carbohydrates in Soil. *Developments in Soil Science* **1978,** *8*, 65-93.

7. Rillig, M. C.; Caldwell, B. A.; Wösten, H. A. B.; Sollins, P., Role of proteins in soil carbon and nitrogen storage: controls on persistence. *Biogeochemistry* **2007,** *85* (1), 25-44.

8. Zelles, L.; Bai, Q. Y., Fractionation of fatty acids derived from soil lipids by solid phase extraction and their quantitative analysis by GC-MS. *Soil Biology and Biochemistry* **1993,** *25* (4), 495-507.

9. Freudenberg, K.; Neish, A. C., *Constitution and biosynthesis of lignin*. Springer-Verlag: Berlin, 1968; p 129 pp.

10. Adapa, P. K.; Karunakaran, C.; Tabil, L. G.; Schoenau, G. J., Potential applications of infrared and Raman spectromicroscopy for agricultural biomass. *Agricultural Engineering International: CIGR Journal* **2009**.

11. MacCarthy, P., The principles of humic substances. *Soil Science* **2001,** *166* (11), 738-751.

12. Simpson, A. J.; Simpson, M. J.; Soong, R., Nuclear Magnetic Resonance Spectroscopy and Its Key Role in Environmental Research. *Environmental Science & Technology* **2012,** *46* (21), 11488-11496.

13. Kononova, M. a. M., Soil organic matter, its nature, its role in soil formation and in soil fertility. *Soil organic matter, its nature, its role in soil formation and in soil fertility.* **1961**.

14. Stevenson, F. J., *Humus chemistry: genesis, composition, reactions*. John Wiley & Sons: 1994.

15. Piccolo, A., The supramolecular structure of humic substances. *Soil science* **2001,** *166* (11), 810-832.

16. Cardoza, L. A.; Korir, A. K.; Otto, W. H.; Wurrey, C. J.; Larive, C. K., Applications of NMR spectroscopy in environmental science. *Progress in nuclear magnetic resonance spectroscopy* **2004,** *45* (3), 209-238.

17. Hertkorn, N.; Kettrup, A., Molecular Level Structural Analysis of Natural Organic Matter and of Humic Substances by Multinuclear and Higher Dimensional NMR Spectroscopy. In *Use of Humic Substances to Remediate Polluted Environments: From Theory to Practice: Proceedings of the NATO Advanced Research Workshop on Use of Humates to Remediate Polluted Environments: From Theory to Practice Zvenigorod, Russia 23–29 September 2002*, Perminova, I. V.; Hatfield, K.; Hertkorn, N., Eds. Springer Netherlands: Dordrecht, 2005; pp 391-435.

18. Remucal, C. K.; Cory, R. M.; Sander, M.; McNeill, K., Low Molecular Weight Components in an Aquatic Humic Substance As Characterized by Membrane Dialysis and Orbitrap Mass Spectrometry. *Environmental Science & Technology* **2012,** *46* (17), 9350-9359.

19. Thomas, J. D., The role of dissolved organic matter, particularly free amino acids and humic substances, in freshwater ecosystems. *Freshwater Biology* **1997,** *38* (1), 1-36.

20. Piccolo, A., *Humic Substances in Terrestrial Ecosystems*. Elsevier Science: 1996.

21. Lal, R.; Follett, R. F.; Kimble, J.; Cole, C. V., Managing U.S. cropland to sequester carbon in soil. *Journal of Soil and Water Conservation* **1999,** *54* (1), 374-381.

22. Lal, R., Soil carbon sequestration to mitigate climate change. *Geoderma* **2004,** *123* (1), 1-22.

23. Spaccini, R.; Piccolo, A.; Conte, P.; Haberhauer, G.; Gerzabek, M. H., Increased soil organic carbon sequestration through hydrophobic protection by humic substances. *Soil Biology and Biochemistry* **2002,** *34* (12), 1839-1851.

24. Tan, K. H., *Humic Matter in Soil and the Environment: Principles and Controversies, Second Edition*. Taylor & Francis: 2014.

25. Mylonas, V. A.; McCants, C. B., Effects of humic and fulvic acids on growth of tobacco. 2. Tobacco growth and ion uptake. *Journal of Plant Nutrition* **1980,** *2* (3), 377-393.

26. Rauthan, B. S.; Schnitzer, M., Effects of a soil fulvic acid on the growth and nutrient content of cucumber (Cucumis sativus) plants. *Plant and Soil* **1981,** *63* (3), 491-495.

27. Adani, F.; Genevini, P.; Zaccheo, P.; Zocchi, G., The effect of commercial humic acid on tomato plant growth and mineral nutrition. *Journal of Plant Nutrition* **1998,** *21* (3), 561-575.

28. Martinez-Balmori, D.; Olivares, F. L.; Spaccini, R.; Aguiar, K. P.; Araújo, M. F.; Aguiar, N. O.; Guridi, F.; Canellas, L. P., Molecular characteristics of vermicompost and their relationship to preservation of inoculated nitrogen-fixing bacteria. *Journal of Analytical and Applied Pyrolysis* **2013,** *104*, 540-550.

29. Canellas, L. P.; Balmori, D. M.; Médici, L. O.; Aguiar, N. O.; Campostrini, E.; Rosa, R. C. C.; Façanha, A. R.; Olivares, F. L., A combination of humic substances and Herbaspirillum seropedicae inoculation enhances the growth of maize (Zea mays L.). *Plant and Soil* **2013,** *366* (1), 119-132.

30. Sutton, R.; Sposito, G., Molecular structure in soil humic substances: the new view. *Environmental science & technology* **2005,** *39* (23), 9009-9015.

31. Claridge, T. D. W., *High-resolution NMR techniques in organic chemistry*. Elsevier: 2016; Vol. 27.

32. Hore, P., *Nuclear magnetic resonance*. Oxford University Press, USA: 2015.

33. Simpson, A. J.; McNally, D. J.; Simpson, M. J., NMR spectroscopy in environmental research: from molecular interactions to global processes. *Progress in Nuclear Magnetic Resonance Spectroscopy* **2011,** *58* (3), 97-175.

34. Hertkorn, N.; Permin, A.; Perminova, I.; Kovalevskii, D.; Yudov, M.; Petrosyan, V.; Kettrup, A., Comparative Analysis of Partial Structures of a Peat Humic and Fulvic Acid Using One- and Two-Dimensional Nuclear Magnetic Resonance Spectroscopy A. Kettrup, present address: Lehrstuhl für Analytische Chemie und Umweltanalytik, Technische Universität München, 85354 Freising, Germany. *Journal of Environmental Quality* **2002,** *31*, 375-387.

35. Ivanova, G.; Randall, E., 13C NMR and mass spectrometry of soil organic matter. In *Open Chemistry*, 2003; Vol. 1, p 10.

36. Simpson, A. J.; Kingery, W. L.; Hatcher, P. G., The Identification of Plant Derived Structures in Humic Materials Using Three-Dimensional NMR Spectroscopy. *Environmental Science & Technology* **2003,** *37* (2), 337-342.

37. Bell, N. G. A.; Michalchuk, A. A. L.; Blackburn, J. W. T.; Graham, M. C.; Uhrín, D., Isotope-Filtered 4D NMR Spectroscopy for Structure Determination of Humic Substances. *Angewandte Chemie (International Ed. in English)* **2015,** *54* (29), 8382-8385.

38. Comisarow, M. B.; Marshall, A. G., Fourier transform ion cyclotron resonance spectroscopy. *Chemical Physics Letters* **1974,** *25* (2), 282-283.

39. Qi, Y.; O'Connor Peter, B., Data processing in Fourier transform ion cyclotron resonance mass spectrometry. *Mass Spectrometry Reviews* **2014,** *33* (5), 333-352.

40. Amster, I. J., Fourier Transform Mass Spectrometry. *Journal of Mass Spectrometry* **1996,** *31* (12), 1325-1337.

41. Wolff, J. J.; Amster, I. J., *Fourier Transform Ion Cyclotron Resonance and MAgnetic Sector Analyzers for ESI and MALDI*. Wiley Books: 2012.

42. Wilson, C. T. R.; Taylor, G. I., The bursting of soap-bubbles in a uniform electric field. *Mathematical Proceedings of the Cambridge Philosophical Society* **2008,** *22* (5), 728-730.

43. Banerjee, S.; Mazumdar, S., Electrospray Ionization Mass Spectrometry: A Technique to Access the Information beyond the Molecular Weight of the Analyte. *International Journal of Analytical Chemistry* **2012,** *2012*, 40.

44. Hautreux, M.; Hue, N.; Du Fou de Kerdaniel, A.; Zahir, A.; Malec, V.; Laprévote, O., Under non-denaturing solvent conditions, the mean charge state of a multiply charged protein ion formed by electrospray is linearly correlated with the macromolecular surface. *International Journal of Mass Spectrometry* **2004,** *231* (2), 131-137.

45. Iribarne, J. V.; Thomson, B. A., On the evaporation of small ions from charged droplets. *The Journal of Chemical Physics* **1976,** *64* (6), 2287-2294.

46. Karas, M.; Bachmann, D.; Bahr, U.; Hillenkamp, F., Matrix-assisted ultraviolet laser desorption of non-volatile compounds. *International Journal of Mass Spectrometry and Ion Processes* **1987,** *78*, 53-68.

47. Zenobi, R.; Knochenmuss, R., Ion formation in MALDI mass spectrometry. *Mass Spectrometry Reviews* **1998,** *17* (5), 337-366.

48. Karas, M.; Krüger, R., Ion Formation in MALDI: The Cluster Ionization Mechanism. *Chemical Reviews* **2003,** *103* (2), 427-440.

49. Dreisewerd, K., The Desorption Process in MALDI. *Chemical Reviews* **2003,** *103* (2), 395-426.

50. de Hoffmann, E.; Stroobant, V., *Mass Spectrometry: Principles and Applications*. Wiley: 2007.

51. Knochenmuss, R., Ion formation mechanisms in UV-MALDI. *Analyst* **2006,** *131* (9), 966-986.

52. Wang, R.-Q.; Druckenmüller, K.; Elbers, G.; Guenther, K.; Croué, J.-P., Analysis of aquatic-phase natural organic matter by optimized LDI-MS method. *Journal of Mass Spectrometry* **2014,** *49* (2), 154-160.

53. Mugo, S. M.; Bottaro, C. S., Characterization of humic substances by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry* **2004,** *18* (20), 2375-2382.

54. Robb, D. B.; Covey, T. R.; Bruins, A. P., Atmospheric Pressure Photoionization: An Ionization Method for Liquid Chromatography−Mass Spectrometry. *Analytical Chemistry* **2000,** *72* (15), 3653-3659.

55. Kauppila, T. J.; Kotiaho, T.; Kostiainen, R.; Bruins, A. P., Negative ion-atmospheric pressure photoionization-mass spectrometry. *Journal of the American Society for Mass Spectrometry* **2004,** *15* (2), 203-211.

56. French, J. B.; Thomson, B. A.; Davidson, W. R.; Reid, N. M.; Buckley, J. A., Atmospheric Pressure Chemical Ionization Mass Spectrometry. In *Mass Spectrometry in Environmental Sciences*, Karasek, F. W.; Hutzinger, O.; Safe, S., Eds. Springer US: Boston, MA, 1985; pp 101-121.

57. Cambier, V.; Hance, T.; de Hoffmann, E., Non-injured maize contains several 1,4-benzoxazin-3-one related compounds but only as glucoconjugates. *Phytochemical Analysis* **1999,** *10* (3), 119-126.

58. Van Krevelen, D. W., Graphical-statistical method for the study of structure and reaction processes of coal. *Fuel* **1950,** *29*, 269-284.

59. Minor, E. C.; Swenson, M. M.; Mattson, B. M.; Oyler, A. R., Structural characterization of dissolved organic matter: a review of current techniques for isolation and analysis. *Environmental Science: Processes & Impacts* **2014,** *16* (9), 2064-2079.

60. Kim, S.; Kramer, R. W.; Hatcher, P. G., Graphical Method for Analysis of Ultrahigh-Resolution Broadband Mass Spectra of Natural Organic Matter, the Van Krevelen Diagram. *Analytical Chemistry* **2003,** *75* (20), 5336-5344.

61. Sleighter, R. L.; Hatcher, P. G., Molecular characterization of dissolved organic matter (DOM) along a river to ocean transect of the lower Chesapeake Bay by ultrahigh resolution electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Marine Chemistry* **2008,** *110* (3), 140-152.

62. Koch, B. P.; Dittmar, T., From mass to structure: an aromaticity index for high-resolution mass data of natural organic matter. *Rapid Communications in Mass Spectrometry* **2006,** *20* (5), 926-932.

63. Mopper, K.; Stubbins, A.; Ritchie, J. D.; Bialk, H. M.; Hatcher, P. G., Advanced Instrumental Approaches for Characterization of Marine Dissolved Organic Matter: Extraction Techniques, Mass Spectrometry, and Nuclear Magnetic Resonance Spectroscopy. *Chemical Reviews* **2007,** *107* (2), 419-442.

64. Hertkorn, N.; Frommberger, M.; Witt, M.; Koch, B. P.; Schmitt-Kopplin, P.; Perdue, E. M., Natural Organic Matter and the Event Horizon of Mass Spectrometry. *Analytical Chemistry* **2008,** *80* (23), 8908-8919.

65. D'Andrilli, J.; Dittmar, T.; Koch, B. P.; Purcell, J. M.; Marshall, A. G.; Cooper, W. T., Comprehensive characterization of marine dissolved organic matter by Fourier transform ion cyclotron resonance mass spectrometry with electrospray and atmospheric pressure photoionization. *Rapid Communications in Mass Spectrometry* **2010,** *24* (5), 643-650.

66. Cao, D.; Huang, H.; Hu, M.; Cui, L.; Geng, F.; Rao, Z.; Niu, H.; Cai, Y.; Kang, Y., Comprehensive characterization of natural organic matter by MALDI- and ESI-Fourier transform ion cyclotron resonance mass spectrometry. *Analytica Chimica Acta* **2015,** *866*, 48-58.

67. Cho, Y.; Jin, J. M.; Witt, M.; Birdwell, J. E.; Na, J.-G.; Roh, N.-S.; Kim, S., Comparing Laser Desorption Ionization and Atmospheric Pressure Photoionization Coupled to Fourier Transform Ion Cyclotron Resonance Mass Spectrometry To

Characterize Shale Oils at the Molecular Level. *Energy & Fuels* **2013,** *27* (4), 1830-1837.

68. Carré, V.; Schramm, S.; Aubriet, F., Study of a complex environmental mixture by electrospray ionization and laser desorption ionization high resolution mass spectrometry: the cigarette smoke aerosol. *AIMS Environmental Science* **2015,** *2* (3), 547-564.

69. Ghosh, K.; Schnitzer, M., Macromolecular Structures Of Humic Substances. *Soil Science* **1980,** *129* (5).

70. Fievre, A.; Solouki, T.; Marshall, A. G.; Cooper, W. T., High-Resolution Fourier Transform Ion Cyclotron Resonance Mass Spectrometry of Humic and Fulvic Acids by Laser Desorption/Ionization and Electrospray Ionization. *Energy & Fuels* **1997,** *11* (3), 554-560.

71. Stenson, A. C.; Landing, W. M.; Marshall, A. G.; Cooper, W. T., Ionization and Fragmentation of Humic Substances in Electrospray Ionization Fourier Transform-Ion Cyclotron Resonance Mass Spectrometry. *Analytical Chemistry* **2002,** *74* (17), 4397-4409.

72. Kujawinski, E. B., Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (ESI FT-ICR MS): Characterization of Complex Environmental Mixtures. *Environmental Forensics* **2002,** *3* (3-4), 207-216.

73. D'Andrilli, J.; Foreman, C. M.; Marshall, A. G.; McKnight, D. M., Characterization of IHSS Pony Lake fulvic acid dissolved organic matter by electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry and fluorescence spectroscopy. *Organic Geochemistry* **2013,** *65*, 19-28.

74. Hertkorn, N.; Benner, R.; Frommberger, M.; Schmitt-Kopplin, P.; Witt, M.; Kaiser, K.; Kettrup, A.; Hedges, J. I., Characterization of a major refractory component of marine dissolved organic matter. *Geochimica et Cosmochimica Acta* **2006,** *70* (12), 2990-3010.

75. Koch, B. P.; Witt, M.; Engbrodt, R.; Dittmar, T.; Kattner, G., Molecular formulae of marine and terrigenous dissolved organic matter detected by electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Geochimica et Cosmochimica Acta* **2005,** *69* (13), 3299-3308.

76. Kujawinski, E. B.; Del Vecchio, R.; Blough, N. V.; Klein, G. C.; Marshall, A. G., Probing molecular-level transformations of dissolved organic matter: insights on photochemical degradation and protozoan modification of DOM from electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Marine Chemistry* **2004,** *92* (1), 23-37.

77. Kellerman, A. M.; Kothawala, D. N.; Dittmar, T.; Tranvik, L. J., Persistence of dissolved organic matter in lakes related to its molecular characteristics. *Nature Geosci* **2015,** *8* (6), 454-457.

78. de Koning, L. J.; Nibbering, N. M. M.; van Orden, S. L.; Laukien, F. H., Mass selection of ions in a Fourier transform ion cyclotron resonance trap using correlated

harmonic excitation fields (CHEF). *International Journal of Mass Spectrometry and Ion Processes* **1997,** *165*, 209-219.

79. Witt, M.; Fuchser, J.; Koch, B. P., Fragmentation Studies of Fulvic Acids Using Collision Induced Dissociation Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Analytical Chemistry* **2009,** *81* (7), 2688-2694.

80. Marshall, A. G.; Lin Wang, T.-C.; Lebatuan Ricca, T., Ion cyclotron resonance excitatio/de-excitation: A basis for Stochastic fourier transform ion cyclotron mass spectrometry. *Chemical Physics Letters* **1984,** *105* (2), 233-236.

81. Ross Iii, C. W.; Guan, S.; Grosshans, P. B.; Ricca, T. L.; Marshall, A. G., Two-dimensional Fourier transform ion cyclotron resonance mass spectrometry/mass spectrometry with stored-waveform ion radius modulation. *Journal of the American Chemical Society* **1993,** *115* (17), 7854-7861.

82. van Agthoven, M. A.; Delsuc, M.-A.; Bodenhausen, G.; Rolando, C., Towards analytically useful two-dimensional Fourier transform ion cyclotron resonance mass spectrometry. *Analytical and Bioanalytical Chemistry* **2013,** *405* (1), 51-61.

83. Ricca, G.; Severini, F.; Di Silvestro, G.; Yuan, C. M.; Adani, F., Derivatization and structural studies by spectroscopic methods of humic acids from Leonardite. *Geoderma* **2000,** *98* (3), 115-125.

84. Wershaw, R. L.; Pinckney, D. J., Methylation of Humic Acid Fractions. *Science* **1978,** *199* (4331), 906.

85. Thorn, K. A.; Steelink, C.; Wershaw, R. L., Methylation patterns of aquatic humic substances determined by13C NMR spectroscopy. *Organic Geochemistry* **1987,** *11* (3), 123-137.

86. Gonzalez-Vila, F. J.; Lüdemann, H. D.; Martin, F., 3C-NMR structural features of soil humic acids and their methylated, hydrolyzed and extracted derivatives. *Geoderma* **1983,** *31* (1), 3-15.

87. Mikita, M. A.; Steelink, C.; Wershaw, R. L., Carbon-13 enriched nuclear magnetic resonance method for the determination of hydroxyl functionality in humic substances. *Analytical Chemistry* **1981,** *53* (11), 1715-1717.

88. Bell, N. G. A.; Murray, L.; Graham, M. C.; Uhrin, D., NMR methodology for complex mixture 'separation'. *Chemical Communications* **2014,** *50* (14), 1694-1697.

89. Kostyukevich, Y.; Kononikhin, A.; Popov, I.; Kharybin, O.; Perminova, I.; Konstantinov, A.; Nikolaev, E., Enumeration of Labile Hydrogens in Natural Organic Matter by Use of Hydrogen/Deuterium Exchange Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Analytical Chemistry* **2013,** *85* (22), 11007-11013.

90. Aoyama, T.; Shioiri, T., Trimethylsilyldiazomethane : a convenient reagent for the o-methylation of alcohols. *Tetrahedron Letters* **1990,** *31* (38), 5507-5508.

91. Kühnel, E.; Laffan, D. D. P.; Lloyd-Jones, G. C.; Martínez del Campo, T.; Shepperson, I. R.; Slaughter, J. L., Mechanism of Methyl Esterification of Carboxylic

Acids by Trimethylsilyldiazomethane. *Angewandte Chemie International Edition* **2007,** *46* (37), 7075-7078.

92. Kew, W.; Blackburn, J. W. T.; Clarke, D. J.; Uhrín, D., Interactive van Krevelen diagrams – Advanced visualisation of mass spectrometry data of complex mixtures. *Rapid Communications in Mass Spectrometry* **2017,** *31* (7), 658-662.

93. Hughey, C. A.; Hendrickson, C. L.; Rodgers, R. P.; Marshall, A. G.; Qian, K., Kendrick Mass Defect Spectrum: A Compact Visual Analysis for Ultrahigh-Resolution Broadband Mass Spectra. *Analytical Chemistry* **2001,** *73* (19), 4676-4681.

94. Kilgour, D. P. A.; Mackay, C. L.; Langridge-Smith, P. R. R.; O'Connor, P. B., Appropriate Degree of Trust: Deriving Confidence Metrics for Automatic Peak Assignment in High-Resolution Mass Spectrometry. *Analytical Chemistry* **2012,** *84* (17), 7431-7435.

95. Konermann, L.; Ahadi, E.; Rodriguez, A. D.; Vahidi, S., Unraveling the Mechanism of Electrospray Ionization. *Analytical Chemistry* **2013,** *85* (1), 2-9.

96. Lewis, J. K.; Wei, J.; Siuzdak, G., Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry in Peptide and Protein Analysis. In *Encyclopedia of Analytical Chemistry*, John Wiley & Sons, Ltd: 2006.

97. Blackburn, J. W. T.; Kew, W.; Graham, M. C.; Uhrín, D., Laser Desorption/Ionization Coupled to FTICR Mass Spectrometry for Studies of Natural Organic Matter. *Analytical Chemistry* **2017,** *89* (8), 4382-4386.

98. Kujawinski, E. B.; Hatcher, P. G.; Freitas, M. A., High-Resolution Fourier Transform Ion Cyclotron Resonance Mass Spectrometry of Humic and Fulvic Acids: Improvements and Comparisons. *Analytical Chemistry* **2002,** *74* (2), 413-419.

99. Shaw, J. B.; Lin, T.-Y.; Leach, F. E.; Tolmachev, A. V.; Tolić, N.; Robinson, E. W.; Koppenaal, D. W.; Paša-Tolić, L., 21 Tesla Fourier Transform Ion Cyclotron Resonance Mass Spectrometer Greatly Expands Mass Spectrometry Toolbox. *Journal of The American Society for Mass Spectrometry* **2016,** *27* (12), 1929-1936.

100. Kido Soule, M. C.; Longnecker, K.; Giovannoni, S. J.; Kujawinski, E. B., Impact of instrument and experiment parameters on reproducibility of ultrahigh resolution ESI FT-ICR mass spectra of natural organic matter. *Organic Geochemistry* **2010,** *41* (8), 725-733.

101. Sleighter, R. L.; Chen, H.; Wozniak, A. S.; Willoughby, A. S.; Caricasole, P.; Hatcher, P. G., Establishing a Measure of Reproducibility of Ultrahigh-Resolution Mass Spectra for Complex Mixtures of Natural Organic Matter. *Analytical Chemistry* **2012,** *84* (21), 9184-9191.

102. Kew, W.; Blackburn, J. W. T.; Uhrín, D., Response to Comment on "Laser Desorption/Ionization Coupled to FTICR Mass Spectrometry for Studies of Natural Organic Matter". *Analytical Chemistry* **2018,** *90* (9), 5968-5971.

103. Stenson, A. C.; Marshall, A. G.; Cooper, W. T., Exact Masses and Chemical Formulas of Individual Suwannee River Fulvic Acids from Ultrahigh Resolution

Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectra. *Analytical Chemistry* **2003,** *75* (6), 1275-1284.

104. Reemtsma, T.; These, A.; Linscheid, M.; Leenheer, J.; Spitzy, A., Molecular and Structural Characterization of Dissolved Organic Matter from the Deep Ocean by FTICR-MS, Including Hydrophilic Nitrogenous Organic Molecules. *Environmental Science & Technology* **2008,** *42* (5), 1430-1437.

105. Cao, D.; Lv, J.; Geng, F.; Rao, Z.; Niu, H.; Shi, Y.; Cai, Y.; Kang, Y., Ion Accumulation Time Dependent Molecular Characterization of Natural Organic Matter Using Electrospray Ionization-Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Analytical Chemistry* **2016,** *88* (24), 12210-12218.

106. Zherebker, A.; Kostyukevich, Y.; Kononikhin, A.; Kharybin, O.; Konstantinov, A. I.; Zaitsev, K. V.; Nikolaev, E.; Perminova, I. V., Enumeration of carboxyl groups carried on individual components of humic systems using deuteromethylation and Fourier transform mass spectrometry. *Analytical and Bioanalytical Chemistry* **2017,** *409* (9), 2477-2488.

Appendix 1 – Python 3 code for automated analysis of CD_3 methylated HS sample

```
import csv
import time
from decimal import *
getcontext().prec = 9

#-----------------------------------------------------------------------
#                     Functions and Array setup
#-----------------------------------------------------------------------

# File reading check function, if file not valid it loops if file
# valid it returns. Aims to trap incorrect inputs such as typing
# errors.
def readfile():
    try:
        the_file = open(thefile, 'r')
        lines = the_file.readlines()
        the_file.close()
        count.append(1)
        return lines
    except IOError as errno:
        print('Incorrect filename. Please re-enter.')

# File printing function, prints 2 columns (spaced by a ' ')
# containing inputs a and b (defined by the information requesting
# to save.
def printfile(a, b, c, d):
    file = open(filename, 'w')
    for i in range(len(a)):
```

```
            x = [a[i], ',', str(b[i]), ',', c[i], ',', d[i], '\r']
            x = ''.join(x)
            x = str(x)
            file.write(x)
        file.close()
        print('Save Complete.')

# Counts number of C.
def Cnumber(x, j):
    ccount = []
    ccount.append(x[j+1])
    if x[j+2] != 'H':
        ccount.append(x[j+2])
    y = ''.join(ccount)
    return y

# Counts number of H.
def Hnumber(x, j):
    hcount = []
    hcount.append(x[j+1])
    if x[j+2] != 'O' and x[j+2] != 'X' and x[j+2] != 'E':
        hcount.append(x[j+2])
        if x[j+3] != 'O' and x[j+3] != 'X' and x[j+3] != 'E':
            hcount.append(x[j+3])
    y = ''.join(hcount)
    return y

# Counts number of O.
def Onumber(x, j):
    ocount = []
    ocount.append(x[j+1])
    if (x[j+2] != 'E') and (x[j+2] != 'X'):
        ocount.append(x[j+2])
    y = ''.join(ocount)
    return y

# Form global arrays
total = []
total1 = []
form = []
form1 = []
mass = []
mass1 = []
stim = []
stim1 = []
CD = []
Ccomp = []
HD = []
HD1 = []
Hcomp = []
OD = []
Ocomp = []
Formula = []
MassCH3 = []
MassCD3 = []
COOHcount = []
Parmass = []

# Error for identifying H3 to D3 shift, allows for calibration
```

```
# discrepancy.
error = Decimal(0.00075)

#------------------------------------------------------------------------
#                              Reading
#------------------------------------------------------------------------

# Asks for molecular file name. Will loop until correct file name is
entered.
count = []
while (len(count) < 1):
    thefile = input('What is the CH3 KMD file name? ')
    lines = readfile()
# Input for deuterated spectra mass list. Again loops.
count = []
while (len(count) < 1):
    thefile = input('What is the CD3 peak list file name? ')
    lines1 = readfile()




#------------------------------------------------------------------------
#                              Processing
#------------------------------------------------------------------------

# Extract formula, formula mass and stimulation from file and then
# remove title line.(INPUT 1)
reader = csv.reader(lines)
for row in reader:
    total.append(row[0])
    total.append(row[1])
    total.append(row[8])

# Extract formula, formula mass and stimulation from file and then
# remove title line.(INPUT 2)
reader = csv.reader(lines1)
for row in reader:
    total1.append(row[0])
total1.remove('m/z')

for i in range(len(total1)):
    y = Decimal(total1[i])
    mass1.append(y)

# Set threshold limit for stimulation.
thresh = input('Set stimulation threshold as? ')
# Separate formula, formula mass and stimulation into separate
# lists, removing any points either below the stimulation threshold
# or that were unassigned by KMD. (Increases the masses to match the
# H returned instead of the negative charge). (INPUT 1)
for i in range(2, len(total), 3):
    if int(total[i]) >= int(thresh) and float(total[i-1]) != 0:
        form.append(total[i-2])
        stim.append(total[i])
        y = Decimal(total[i-1])
        mass.append(y)
```

```
# Section to scan through formula and masses of CH3 methylated list,
# and search for multiples of 3D mass increases (within small error)
# for 1 to #O/2 (rounded down) COOH groups. Converts all values into
# strings.
for i in range(len(form)):
    lower = 0
    upper = 0
    x = list(form[i])
    x.append('E')
    for j in range(len(x)):
        if x[j] == 'O':
            z = Onumber(x, j)
            z = int(z)
            COOHlimit = z/2
            COOHlimit = int(COOHlimit)
            y = Hnumber(x, j)
            y = int(y)
            Hlimit = y/3
            Hlimit = int(Hlimit)
            if Hlimit < COOHlimit:
                COOHlimit = Hlimit
            for k in range(1, (COOHlimit+1)):
                lower = mass[i] + Decimal(k*3.01883) - error
                upper = mass[i] + Decimal(k*3.01883) + error
                for l in range(len(mass1)):
                    if (mass1[l]>=mass[i]) and
(mass1[l]<(mass[i]+31)):
                        if mass1[l]>=lower and mass1[l]<upper:
                            Formula.append(form[i])
                            MassCH3.append(mass[i])
                            MassCD3.append(mass1[l])
                            COOHcount.append(k)
                            break
for i in range(len(MassCH3)):
    MassCH3[i] = str(MassCH3[i])
    MassCD3[i] = str(MassCD3[i])
    COOHcount[i] = str(COOHcount[i])

# Takes all formula, removes negative charge and adds 1H to the
# formula. Then corrects the formula for by -1C and -2H for each
# COOH group in the formula.
for i in range(len(Formula)):
    hcount = []
    newform = []
    marker = 0
    y = list(Formula[i])
    for j in range(len(y)):
        if y[j] == 'C':
            marker = 1
            newform.append(y[j])
            z = Cnumber(y, j)
            z = int(z) - int(COOHcount[i])
            newform.append(str(z))
        if y[j] == 'H':
            marker = 1
            newform.append(y[j])
            z = Hnumber(y, j)
            z = int(z) - (int(COOHcount[i])*2)
```

```
            newform.append(str(z))
        if y[j] == 'O':
            marker = 0
        if y[j] == 'E':
            break
        elif marker != 1:
            newform.append(y[j])
    a = ''.join(newform)
    Formula[i] = a

# Calculates the formula mass of the parent compound for convenient
# graphical analysis.
for i in range(len(Formula)):
    m = 0
    y = list(Formula[i])
    y.append('E')
    for j in range(len(y)):
        if y[j] == 'C':
            z = Cnumber(y, j)
            z = Decimal(int(z)*12.0000)
            m = m + z
        if y[j] == 'H':
            z = Hnumber(y, j)
            z = Decimal(int(z) * 1.007825)
            m = m + z
        if y[j] == 'O':
            z = Onumber(y, j)
            z = int(z)
            z = Decimal(z*15.994915)
            m = m + z
    Parmass.append(m)

#-----------------------------------------------------------------------
#                                  Writing
#-----------------------------------------------------------------------

# Writes entire data file, in format suitable for a .txt file,
# currently as columns of the 2 data types chosen (formula and
# mass).
print('Data suitable for .txt or .csv files.')
filename = input('Save COOH list file as: ')
printfile(Formula, Parmass, MassCD3, COOHcount)
```